EpiRR
A Registry for Reference Epigenomes

Alessandro Vullo on behalf of

David Richardson
Avik Datta
Laura Clarke
Paul Flicek

EMBL – European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK

Background

- IHEC goal

  "produce and archive >1000 reference epigenomes"

- Ensure generated data is useful for comparative/integrative analyses by others
- Distributed data production
  - challenging!

IHEC Data

Assay type	Required	Optional
DNA methylation	BS-seq	MDB, MeDIP, MRE, RRBS
Histone modification	Input, H3K27me3, H3K36me3, H3K4me3, H3K27ac, H3K9me3	Any other histone modifications
RNA-seq	mRNA	miRNA, smRNA
Chromatin accessibility		FAIRE-seq, DnaseI-seq, ATAC-seq

Problem

- Public archives metadata do not facilitate linking different experiments
- Biological material used and location of data must be known
- Need a tracking system

EpiRR

http://www.ebi.ac.uk/vg/epirr

- Holds public archive accessions
- Provides accessions for reference epigenomes
  - discrete, coherent units

Data Model

- A reference epigenome
  - belongs to a member project
  - receives an accession when created
- Assumed will be submitted before completed
  - updates are supported
  - versioning to ensure precise identification and metadata updates

Submission Workflow

- Accepts text/JSON
- Validates metadata associated to component data sets
  - error if not enough descriptive
  - experiment used to assess completeness
- Assigns unique identifiers to reference epigenomes
- Summarises key experimental characteristics
  - e.g. single or pool of samples

Availability

- Minimal web interface

  http://www.ebi.ac.uk/vg/epirr

- REST API
  - **/view/all** – returns all current datasets
  - **/view/:id** – returns single dataset
  - **/summary** – dataset counts by project

/view/all

BLUEPRINT	Single donor	Incomplete	IHECRE00000345.1		segmented neutrophil of bone marrow from Bone marrow of BM190913
BLUEPRINT	Single donor	Incomplete	IHECRE00000346.1		mature neutrophil from Venous blood of PB130513
BLUEPRINT	Single donor	Incomplete	IHECRE00000347.1		plasma cell from Tonsil of T14/10
BLUEPRINT	Single donor	Incomplete	IHECRE00000348.1		Acute Myeloid Leukemia from Bone marrow of UMCG00012
BLUEPRINT	Single donor	Incomplete	IHECRE00000349.1		naive B cell from Venous blood of NC14/5
BLUEPRINT	Single donor	Incomplete	IHECRE00000350.1		Plasma cell from Tonsil of T12-18
BLUEPRINT	Single donor	Incomplete	IHECRE00000351.1		effector memory CD4-positive, alpha-beta T cell from Venous blood of S001U3
BLUEPRINT	Single donor	Incomplete	IHECRE00000352.1		Multiple myeloma from Bone marrow of 23376
BLUEPRINT	Single donor	Incomplete	IHECRE00000353.1		mature neutrophil - G-CSF/Dex. Treatment (16-20 hrs) from Venous blood of PB100713
BLUEPRINT	Single donor	Incomplete	IHECRE00000354.1		naive B cell from Venous blood of NC11/41
NIH Roadmap Epigenomics	Single donor	Incomplete	IHECRE00000924.2	roadmap-epigenomics:E099	REMC Epigenome (Class 5) for Placenta Amnion using donors/samples:CTL02
NIH Roadmap Epigenomics	Composite	Incomplete	IHECRE00000925.2	roadmap-epigenomics:E089	REMC Epigenome (Class 5) for Fetal Muscle Trunk using donors/samples:H-24851

/view/:id

IHECRE00000003.2

Type	Single donor
Status	Incomplete
Project	BLUEPRINT
Local name	
Description	regulatory T cell from Venous blood of C001FR
Is live version?	yes
Other versions	previous

Metadata

biomaterial_type	Primary Cell
disease_ontology_uri	NA
donor_id	C001FR
passage_if_expanded	NA
donor id	C001FR
donor_ethnicity	Northern European
donor_sex	Female
gender	female
tissue_type	Venous blood
taxon_id	9606
species	Homo sapiens
disease	None
sample_ontology_uri	http://purl.obolibrary.org/obo/CL_0000815;http://purl.obolibrary.org/obo/UBERON_0013756
donor_age	60 - 65
phenotype	CL:0000815;EFO:0000761;UBERON:0013756
markers	CD3+ CD4+ CD25+ CD127low
cell_type	regulatory T cell
molecule	total RNA
biomaterial_provider	NIHR Cambridge BioResource
donor_health_status	NA

Raw data

Assay type	Experiment type	Archive	Primary ID	Secondary ID	Link
RNA-Seq	mRNA-seq	EGA	EGAX00001071827	EGAD00001001546	View in archive

/summary

EpiRR dataset summary

Project	Complete	Incomplete	Total dataset count
BLUEPRINT	4	298	302
CEEHRC (CEMT)	48	0	48
CEEHRC (McGill)	13	271	284
NIH Roadmap Epigenomics	23	75	98
Total	88	644	732

Present & Future

- >1400 registered experiments across >300 accessioned epigenomes
- IHEC data portal – full integration
  - display only EpiRR registered data sets
  - tracks required to have an EpiRR accession
- Fall 2016
  - >7000 experiments
  - 1500 reference epigenomes

Acknowledgements

- EBI
  - David Richardson
  - Avik Datta
  - Laura Clarke
  - Paul Flicek
- David Bujold (McGill)
- Sita Gakkhar (BCGSC)

EMBL

IHEC
International Human Epigenome Consortium

BLUEPRINT
epigenome

European Commission

Framework Programme 7