Reproducible big data science: A case study in continuous FAIRness

Kyle Chard

University of Chicago and Argonne National Laboratory

Ravi Madduri, Michael D'Arcy, Segun Jung, Alexis Rodriguez, Dinanath Sulakhe, Eric Deutsch, Cory Funk, Ben Heavner, Matthew Richards, Paul Shannon, Gustavo Glusman, Nathan Price, Carl Kesselman and Ian Foster

Reproducibility requires continuous FAIRness

- Make all data findable, accessible, interoperable, reusable **at every stage**, via pervasive use of simple identifier and exchange format conventions

- Build on proven **security, data, identifier, and computation building blocks** that have large user communities inside and outside biomedicine

- Leverage industry best practices to meet **scalability, interoperability, sustainability, and reliability** needs

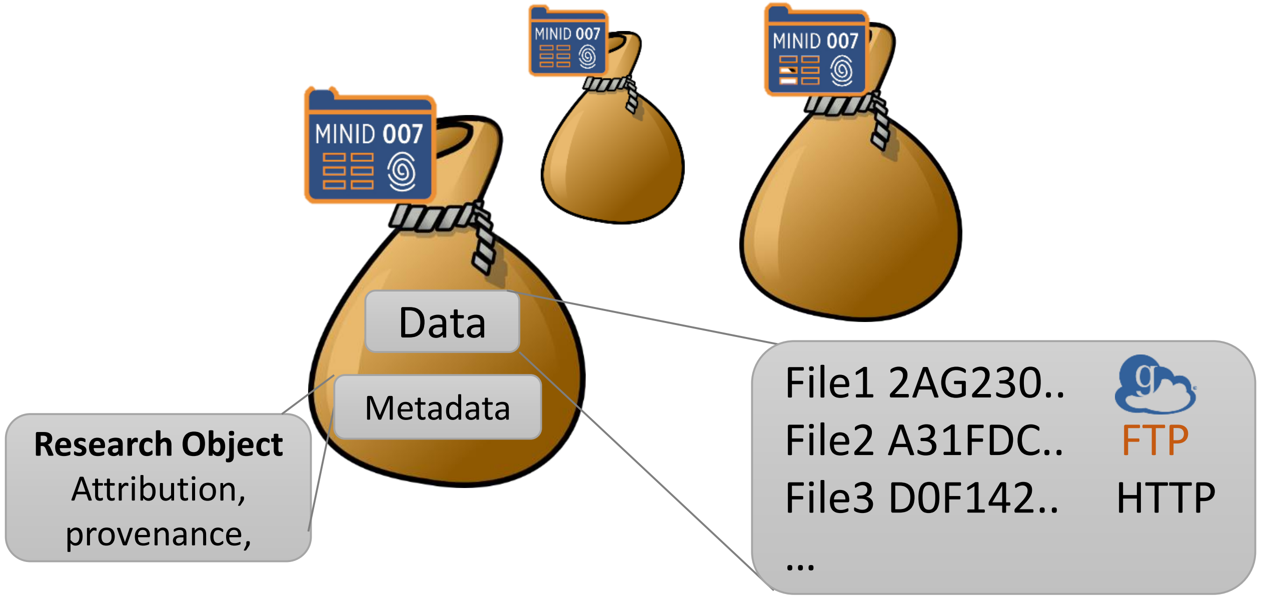
Interoperability: naming and exchange

Minid

- Lightweight identifiers for any product at any stage
- Easily created, dereferenced, validated
- Global integrity – validate content across the commons

BDBag

- Self-describing and flexible format for exchange
- Extended BagIt Specification
- Standard manifest representation that supports different protocols
- Research Object metadata



Data

Metadata

Research Object
Attribution,
provenance,

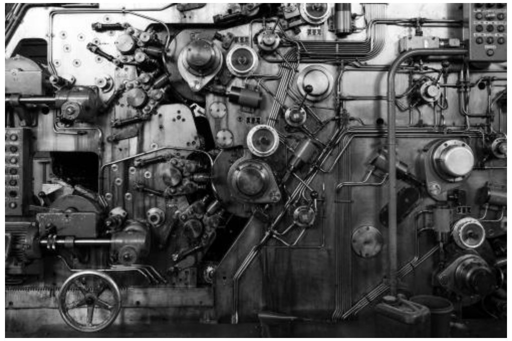
File1 2AG230..
File2 A31FDC..
File3 D0F142..
…

FTP
HTTP

Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data Across 27 Tissue Types

Cory C Funk, Segun Jung, Matthew A Richards, Alex Rodriguez, Paul Shannon, Rory Donovan, Ben Heavner, Kyle Chard, Yukai Xiao, Gustavo Glusman, Nilufer Erteskin-Taner, Todd Golde, Arthur Toga, Leroy Hood, John D Van Horn, Carl Kesselman, Ian Foster, Seth Ament, Ravi Madduri, Nathan D Price

doi: https://doi.org/10.1101/252023

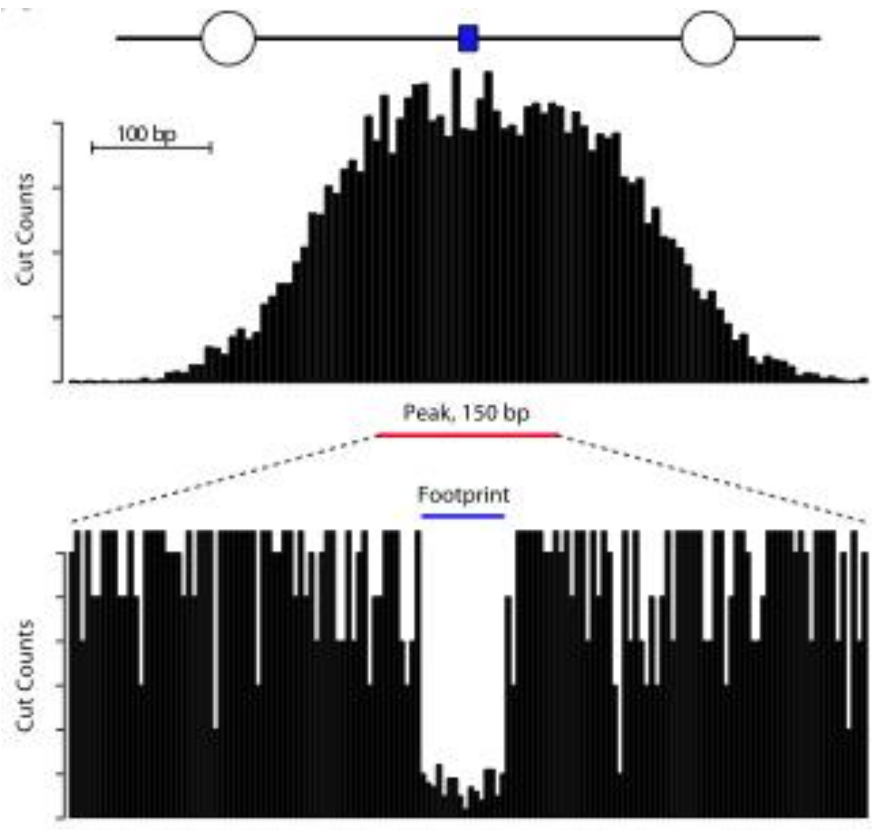


Reproducible big data science: A case study in continuous FAIRness

Ravi K Madduri, Kyle Chard, Mike D'Arcy, Segun C Jung, Alexis Rodriguez, Dinanath Sulakhe, Eric W Deutsch, Cory Funk, Ben Heavner, Matthew Richards, Paul Shannon, Gustavo Glusman, Nathan Price, Carl Kesselman, Ian Foster

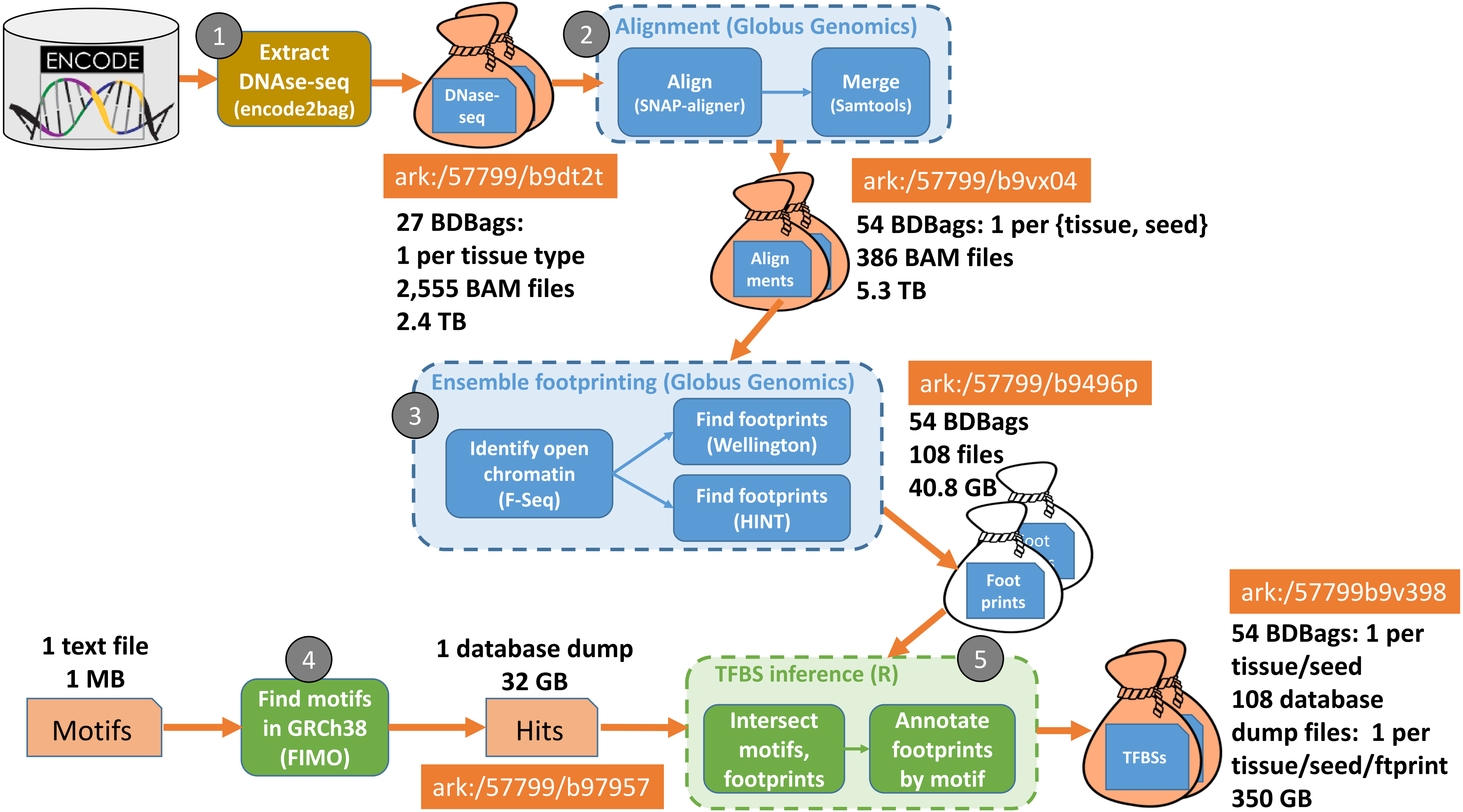
doi: https://doi.org/10.1101/268755

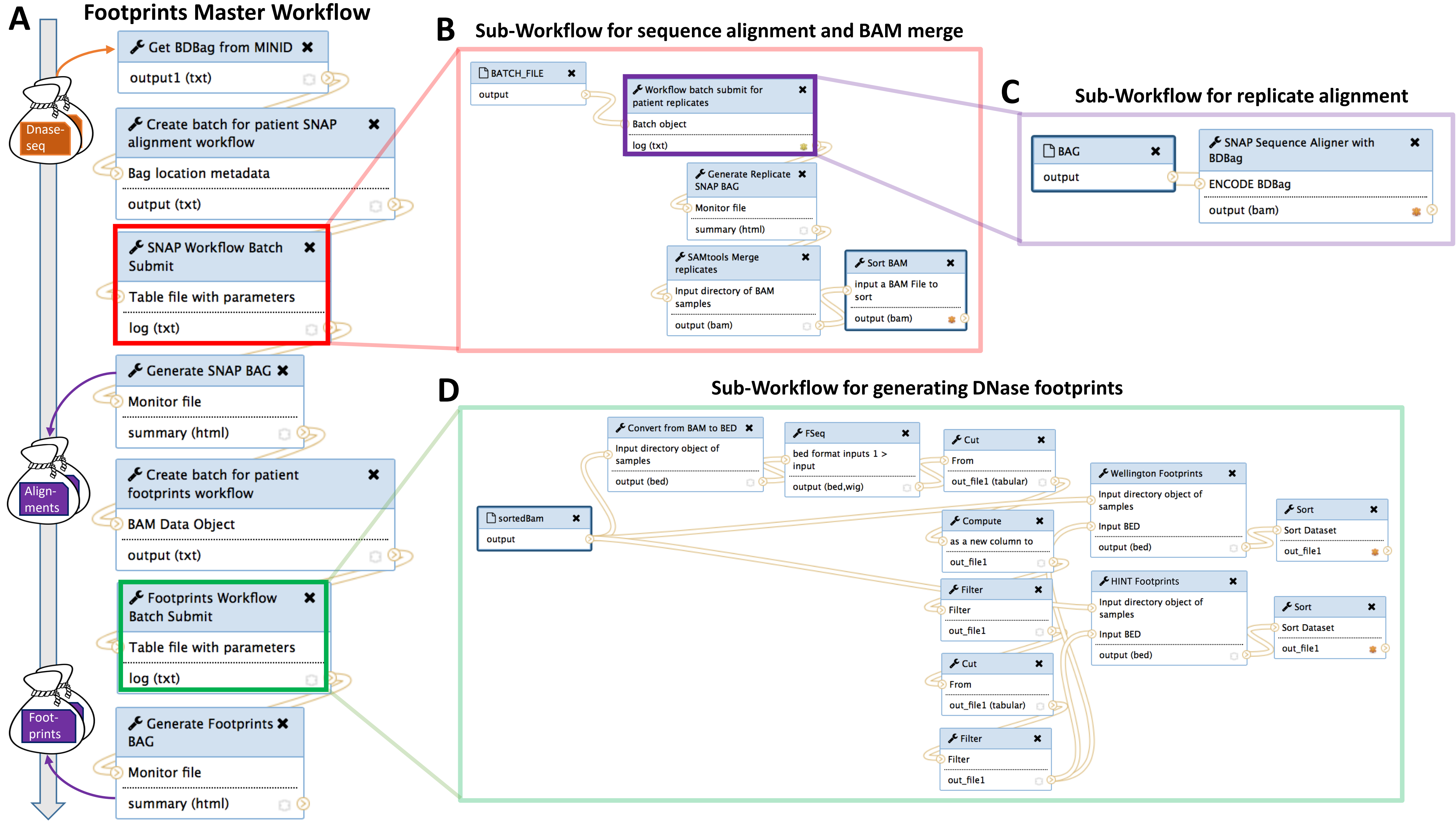
Generation of TFBS Atlas



Vernot et al. doi: 10.1101/gr.134890.111

- Uniform processing of next generation sequencing data
    - Align to reference genome
    - Identify DNase hypersensitve regions
    - Apply multiple footprinting algorithms to locate putative transcription factor binding sites (TFBSs)
- Evaluate confidence in putative TFBSs
- Use TFBSs as features for machine learning approaches applied to disease-specific research


ENCODE
1
Extract DNAse-seq (encode2bag)
DNAse-seq
2
Alignment (Globus Genomics)
Align (SNAP-aligner)
Merge (Samtools)
ark:/57799/b9dt2t
27 BDBags:
1 per tissue type
2,555 BAM files
2.4 TB
Align ments
ark:/57799/b9vx04
54 BDBags: 1 per {tissue, seed}
386 BAM files
5.3 TB
Ensemble footprinting (Globus Genomics)
3
Identify open chromatin (F-Seq)
Find footprints (Wellington)
Find footprints (HINT)
ark:/57799/b9496p
54 BDBags
108 files
40.8 GB
Foot prints
1 text file
1 MB
Motifs
4
Find motifs in GRCh38 (FIMO)
1 database dump
32 GB
Hits
ark:/57799/b97957
TFBS inference (R)
5
Intersect motifs, footprints
Annotate footprints by motif
TFBSs
ark:/57799b9v398
54 BDBags: 1 per tissue/seed
108 database dump files: 1 per tissue/seed/ftprint
350 GB



A Footprints Master Workflow

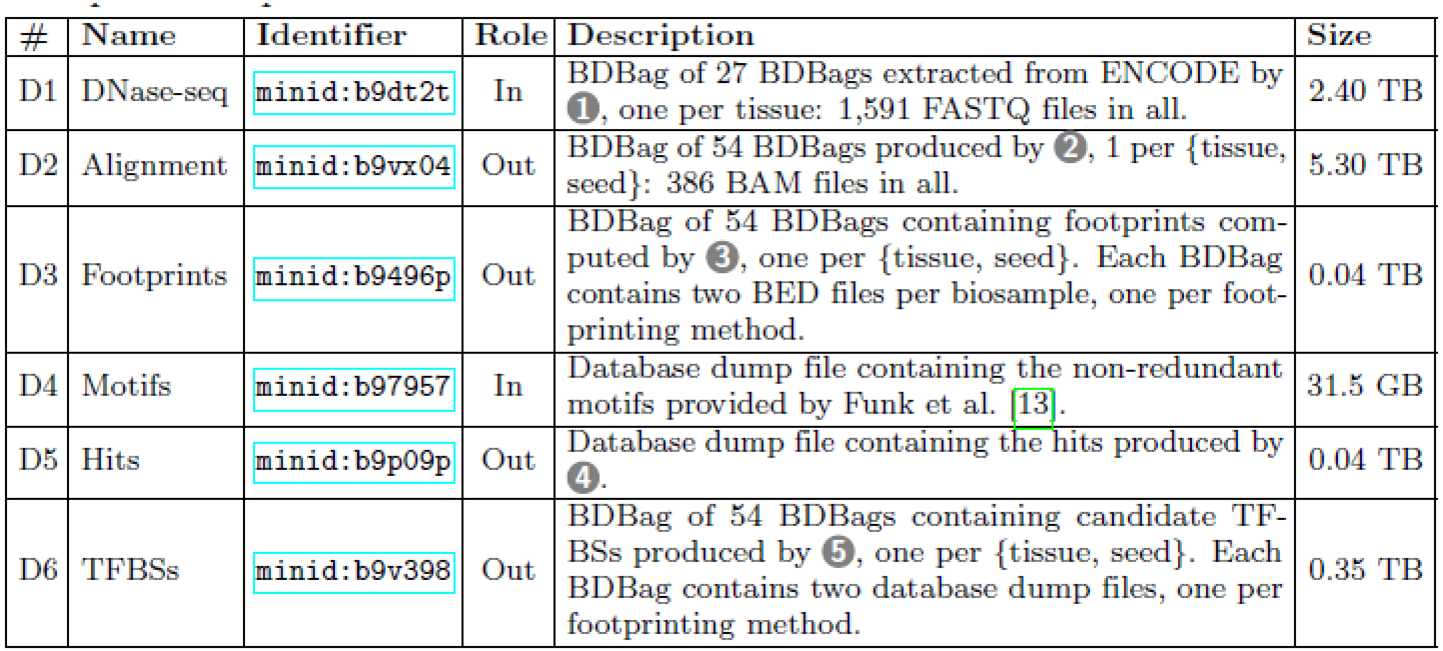
B Sub-Workflow for sequence alignment and BAM merge

C Sub-Workflow for replicate alignment

D Sub-Workflow for generating DNase footprints

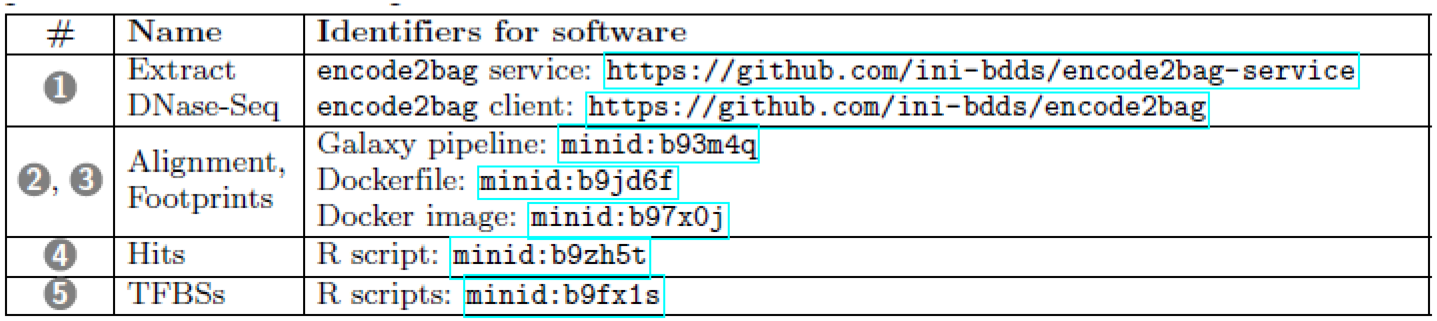
Reproducibility

1) Datasets



#	Name	Identifier	Role	Description	Size
D1	DNase-seq	minid:b9dt2t	In	BDBag of 27 BDBags extracted from ENCODE by ❶, one per tissue: 1,591 FASTQ files in all.	2.40 TB
D2	Alignment	minid:b9vx04	Out	BDBag of 54 BDBags produced by ❷, 1 per {tissue, seed}: 386 BAM files in all.	5.30 TB
D3	Footprints	minid:b9496p	Out	BDBag of 54 BDBags containing footprints computed by ❸, one per {tissue, seed}. Each BDBag contains two BED files per biosample, one per footprinting method.	0.04 TB
D4	Motifs	minid:b97957	In	Database dump file containing the non-redundant motifs provided by Funk et al. [13].	31.5 GB
D5	Hits	minid:b9p09p	Out	Database dump file containing the hits produced by ❹.	0.04 TB
D6	TFBSs	minid:b9v398	Out	BDBag of 54 BDBags containing candidate TFBSs produced by ❺, one per {tissue, seed}. Each BDBag contains two database dump files, one per footprinting method.	0.35 TB

1) Tools



#	Name	Identifiers for software
❶	Extract DNase-Seq	encode2bag service: https://github.com/ini-bdds/encode2bag-service encode2bag client: https://github.com/ini-bdds/encode2bag
❷, ❸	Alignment, Footprints	Galaxy pipeline: minid:b93m4q Dockerfile: minid:b9jd6f Docker image: minid:b97x0j
❹	Hits	R script: minid:b9zh5t
❺	TFBSs	R scripts: minid:b9fx1s

Summary

- Complexity and level of effort increases with size of the data, computation, analysis team
  - TFBS example: big data (~10 TB), parallel computing (70K core hours), distributed data and processing, handful of researchers …
- Continuous FAIRness requires various tools & services
  - Identifiers, BDBags, ROs, Containers,
  - Globus, Globus Genomics, GitHub
- User Study
  - 10/11 students and researchers were able to reproduce this study without assistance