The 4D Nucleome Data Portal: a resource for searching and visualizing curated nucleomics data

Sarah B. Reiff[1], Andrew J. Schroeder[1], Koray Kirli[1], Andrea Cosolo[1], Clara Bakker[1], Luisa Mercado[1], Soohyun Lee[1], Alexander D. Veit[1], Alexander K. Balashov[1], Carl Vitzthum[1], William Ronchetti[1], Kent M. Pitman[1], Jeremy Johnson[1], Shannon R. Ehmsen[1], Peter Kerpedjiev[1], Nezar Abdennur[2], Maxim Imakaev[2], Serkan Utku Öztürk[3], Uğur Çamoğlu[3], Leonid A. Mirny[2], Nils Gehlenborg[1], Burak H. Alver[1], Peter J. Park[1,*]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Department of Physics and Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Karya SMD Software Solutions, İzmir, Turkey

* Corresponding author. Contact: peter_park@hms.harvard.edu

Abstract

The 4D Nucleome (4DN) Network aims to elucidate the complex structure and organization of chromosomes in the nucleus and the impact of their disruption in disease biology. We present the 4DN Data Portal (https://data.4dnucleome.org/), a repository for datasets generated in the 4DN network and relevant external datasets. Datasets were generated with a wide range of experiments, including chromosome conformation capture assays such as Hi-C and other innovative sequencing and microscopy-based assays probing chromosome architecture. All together, the 4DN data portal hosts more than 1800 experiment sets and 34000 files. Results of sequencing-based assays from different laboratories are uniformly processed and quality-controlled. The portal interface allows easy browsing, filtering, and bulk downloads, and the integrated HiGlass genome browser allows interactive visualization and comparison of multiple datasets. The 4DN data portal represents a primary resource for chromosome contact and other nuclear architecture data for the scientific community.

Keywords: 4D Nucleome, Hi-C, chromosome conformation capture, nucleomics, genomics, chromatin, nuclear architecture.

Background

The 4D Nucleome (4DN) Network[1] is an NIH Common Fund project that started in 2015 with the overarching goal of elucidating the three-dimensional organization of chromosomes in the nucleus across different cell types and cell cycle stages, and to understand how perturbation of this structure can impact human health in cases of cancer and other diseases. The first phase (2015-2020) of the consortium included 5 projects on nucleomics technology development, 9 projects on the development of imaging tools, 6 projects studying nuclear bodies and compartments, 6 centers studying nuclear organization, and 10 additional collaborating projects, involving approximately 100 labs in total. The second phase of the consortium (2020-2025) includes 8 projects studying real-time chromatin dynamics, 16 projects examining the role of nuclear biology in human health and disease, and 4 centers developing methods for data integration, modelling, and visualization. The new projects bring the total of 4DN network labs to over 150 and promises the generation of even more nucleomics data, particularly with relevance to human health, as well as more advanced modes of data integration across disparate assay types.

The 4DN Data Coordination and Integration Center (DCIC) has created a web portal that serves as a repository for data generated by the 4DN Network members. It has also imported external datasets that are widely used by the community. Implemented fully on the Amazon Web Services (AWS) platform with the latest technologies, the portal has been engineered to provide utility to the broader community of nuclear biology researchers. It enables easy searching and browsing of the data and, importantly, the associated metadata, thus allowing for increased reproducibility both at the analytical and experimental level.

The portal was designed to accommodate data generated from both genomics and microscopy experiments. A large portion of the genomics data comes from chromosome conformation capture (3C) assays such as Hi-C[2]. Additional data derive from several other genomic assay types that probe chromosome conformation or other aspects of nuclear structure and function, such as replication timing, chromatin accessibility or spatial proximity of DNA to other cellular components. These assays include Repli-seq[3], DamID[4], CUT&RUN[5] and SPRITE[6]. We have incorporated elements specific for imaging modalities into our data model as well, and provide visualization capabilities for some imaging data types. Many current datasets use imaging as a tool for measuring distances between genomic locations or between a genomic region and a sub-nuclear structure, and imaging is often used as a cross-validation technique for results of genomic assays. Current imaging data types available include standard fluorescent in situ hybridization (FISH) that targets a DNA or RNA sequence, multiplexed FISH that can target many loci in a single fixed sample[7], and dynamic single particle tracking.

In addition to making data easily accessible to users, the 4DN DCIC has aimed to develop and run standardized bioinformatics pipelines on submitted raw data to generate consistent and comparable results. These can then be explored in an integrated HiGlass browser that supports visualization of 2-D contact maps and linear 1-D tracks such as gene annotations or CUT&RUN peaks[8]. The portal also supports download of raw and processed files for users that want to use the files locally.

Here we describe the 4D Nucleome Data Portal, a repository for genomic and microscopy nuclear architecture datasets. We discuss data that can be found in the portal, interactive visualization capabilities for these datasets in the portal, and utility to users outside the 4DN network.

4DN Data at the Data Portal

The majority of datasets in the portal are submitted by the 4DN researchers who performed the studies. A subset of datasets on the 4DN data portal were generated outside the consortium but were curated by 4DN curators because they were widely used in the community; these data were uploaded from other public repositories such as NCBI's Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). Within the data model of the 4DN data portal, the top level item is called an Experiment Set, and represents a set of replicate experiments, all performed under the same experimental conditions; as such, quantities of data will be discussed below in terms of experiment sets. 4DN data generators are strongly encouraged to include multiple biological replicates for every experiment, and currently over 800 experiment sets in the portal have met this standard.

Genomics Datasets

For genomics assays, Hi-C and its variants make up the largest proportion of experiments in the 4DN data portal. The original Hi-C protocol was reported in 2009[2], as a chromosome conformation capture (3C) method that assays pairwise interactions in DNA across the whole genome by making use of high-throughput sequencing. Since then, many variants of Hi-C have been described. The first Hi-C experiments used 1% formaldehyde cross-linking, a six-cutter restriction enzyme, and digestion and ligation steps were performed in lysed nuclei in a dilute solution. Later, an *in situ* version involved performing the digestion and ligation steps in intact nuclei. Other types of Hi-C assays involve using different cross-linking chemistries and chromatin fragmentation methods (Micro-C, DNase Hi-C, Hi-C 3.0)[9–11], and single cell and single nucleus versions of Hi-C have also been implemented[12–14]. In total, the 4DN data portal has 597 public Hi-C experiment sets, spanning 9 subtypes of Hi-C assays (Table 1). Some of these datasets are particularly useful because of their high sequencing depth. Over 40 Hi-C experiment sets on the portal have over 1 billion read pairs after filtering, and a table of these is available in the data portal resource pages (https://data.4dnucleome.org/resources/data-collections/high-resolution-hic-datasets).

The same year that Hi-C was reported, the first ChIA-PET (Chromatin Interaction Analysis by Paired End Tag sequencing) experiments were published. Similar to Hi-C in design, ChIA-PET uses chromatin immunoprecipitation to enrich DNA-protein complexes and then employs proximity ligation and sequencing, enabling the discovery of chromatin interactions mediated by a target

protein[15,16]. Together with ChIA-Drop, a droplet-based assay similar to ChIA-PET[17], and PLAC-seq, another similar assay in which proximity ligation is performed prior to immunoprecipitation[18], IP-based 3C assays comprise 32 public experiment sets on the 4DN data portal (Table 1).

Recently, a great deal of exciting technology has been developed by members of the 4DN community, resulting in new innovative assays for probing aspects of nuclear DNA structure. To give a few examples, SPRITE is a method for identifying multi-way interactions between distal genomic regions[6]; MARGI is a method for mapping RNA-chromatin interactions[19–21]; CUT&RUN is a DNA-binding assay generating data similar to ChIP-seq but which produces much lower background[5]; and TSA-seq is being used to analyze the proximity of nuclear structures to genomic regions on a genome-wide scale[22]. All the genomic assays currently on the 4DN data portal and available to the public are shown in (Table 1).

Table 1. Genomic assay types in the 4D Nucleome Data Portal.

Experiment Type	# of Public Experiment Sets
Hi-C	**597**
in situ Hi-C	329
Dilution Hi-C	118
DNase Hi-C	21
Micro-C	26
Single cell Hi-C	11
Single nucleus Hi-C	17
sci-Hi-C[13]	28
Capture Hi-C[28]	40
TCC	7
IP-based 3C assays	**32**
ChIA-PET	4
in situ ChIA-PET	8
ChIA-Drop[17]	2
PLAC-seq[18]	18
DNA Binding Assays	**204**
ChIP-seq	141
CUT&RUN	63
Other Sequencing Assays	**393**
Repli-seq[29]	113
SPRITE[6]	3
DamID	66
ATAC-seq	20
RNA-seq	84
TRIP[30]	7
NAD-seq[31]	8
TSA-seq[22]	67
MARGI[19]	6
GAM[32]	6
RE-seq(DpnII-seq[33])	11
Bru-seq[34]	1
MC-3C[35]	1
Total	**1226**

Genomics Pipelines and QC

For genomics datasets, data are submitted as raw fastq files. All data are aligned to the genome references hg38 and mm10 for analyses of human and mouse samples, respectively. For Hi-C experiments, the general analysis pipeline involves (1) aligning reads to a reference genome; (2) filtering aligned reads; (3) matrix aggregation and normalization; and (4) combining replicates to generate a single contact matrix. At each of these steps, different tools and parameters are employed by independent labs. Although optimal tools/parameters may depend on the type of downstream analysis performed, a uniform processing pipeline was necessary to ensure consistency and compatibility. After extensive discussions and testing by the members of the 4DN Analysis Working Group, the final version of the Hi-C processing included the following steps: (1) alignment by BWA MEM[23] with the -SP5M option to ensure that paired reads are aligned independently but the results are formatted properly as paired-end data and that the 5' portion of a chimeric alignment is reported as a primary soft-clipped alignment, (2) sorting and filtering the reads using pairtools[24] and (3) aggregating filtered reads into a contact matrix and normalizing it. The analysis also includes quality control steps using Fastqc[25] and Pairsqc[26]. This pipeline outputs a multi-resolution contact matrix at the 4DN standard resolutions of 1kb, 2kb, 5kb, 10kb, 25kb, 50kb, 100kb, 250kb, 500kb, 1Mb, 2.5Mb, 5Mb and 10Mb. This matrix is generated using the cooler software and the .mcool file format[27], which is compatible with the HiGlass interactive 2D genome browser[8] and can be visualized on the portal. An additional contact matrix is also generated in .hic format, which is compatible with the Juicebox 2D interactive genome browser[36]. More details of the pipeline can be found in the portal's resource pages at https://data.4dnucleome.org/resources/data-analysis/hi_c-processing-pipeline.

Genomic contact matrices generally show evidence of genomic compartments[2] as well as local regions of enriched intra-compartment contacts, known as topologically-associated domains (TADs)[37,38]. There is a great deal of interest in exploring the nature and dynamics of TADs and sub-TADs, domains nested within

others. Identification of such domains is difficult due to the lack of validated sets, and detection algorithms are continually being refined[39–41]. At this time, the 4DN Data Portal runs two domain identification workflows on contact matrices to report compartments and TAD boundaries, using the cooltools software[42]. The first workflow uses an eigenvector decomposition of the matrix to call active (A) and inactive (B) compartments. The second workflow computes insulation scores along the diagonal of the matrix, based on average interaction frequencies crossing over each genomic bin, and prominent dips in this score indicate boundaries between domains[43]. As the nature of TADs and domains is an area of active investigation, the 4DN DCIC chooses to report boundaries based on insulation scores to provide results that might be further utilized in developing domain detection algorithms. Results of both workflows are accessible as bed files as well as bigwig files which can be visualized in HiGlass on the data portal. More details on the domain-detection pipelines can be found at https://data.4dnucleome.org/resources/data-analysis/insulation_compartment_scores.

Processing pipelines are run on other types of high-throughput sequencing assays at the portal as well, including Repli-seq, CUT&RUN, and MARGI (Table 2). The pipeline for Repli-seq involves (1) trimming reads with cutadapt and aligning to a reference genome with bwa-mem[23]; (2) filtering valid alignments with samtools[44]; and (3) binning and aggregating for 5kb windows with bedtools[45]. Final output is provided in gzipped bedgraph and bigWig formats.

The CUT&RUN pipeline on the data portal processes paired-end reads in three main steps: (1) reads are processed with Trimmomatic for quality filtering and adapter trimming, and aligned to a reference genome with bowtie2[46]; (2) duplicate reads are filtered out using Picard[47] and samtools[44], and converted into .bed format with bedtools[45]; and (3) peaks are called with SEACR[48]. The final outputs of the pipeline include a peaks file and a bigWig track. The final bigWig files of the CUT&RUN and Repli-seq pipelines can both be visualized on the portal as 1D tracks in the HiGlass browser. The MARGI pipeline is similar to the Hi-C pipeline, and is adapted from the original pipeline written by the iMARGI creators[21].

For ATAC-seq, ChIP-seq, and RNA-seq data, pipelines from the ENCODE Data Coordination Center[49] have been adapted for the 4DN platform. For ATAC-seq and ChIP-seq, the final output of these pipelines includes a bigwig file containing the fold change in signal across the genome, which can be visualized as a 1D genome track in HiGlass, and a corresponding quality control report. The peak calling step of these processing pipelines also yields optimal peaks and conservative peaks, which are both available on the portal in bigBed format. For RNA-seq, the final output is a bigWig file which contains read counts, and can be visualized as a 1D track in HiGlass. In addition, .tsv files of gene expression and isoform expression are also available.

Several types of quality control and assessment are performed on submitted and pipeline generated results. FastQC[25] is run on all fastq files, and the report generated is stored in the cloud and viewable to users. PairsQC[26] is a software package developed to assess the quality of .pairs files generated by the Hi-C and MARGI pipelines, and these results are also available on the portal for viewing. Additional QC reports are available on the portal for .bam alignments, as well as ChIP-seq, RNA-seq, ATAC-seq, and Repli-seq results, and report various metrics such as percentage of reads mapped in the case of alignments, or overlap reproducibility measures in the case of ATAC-seq and ChIP-seq.

A primary advantage of using the 4DN Hi-C data is that all similar experiments have been processed in an automated fashion with the same software and versions, so the results are directly comparable and more amenable to meta-analysis. All data are processed by open source software from start (fastq files) to finish (contract matrices and domain calls). All intermediate files and full provenance graphs are available, so that the user can easily find the processing steps and reproduce any portion or the full pipeline. All 4DN pipelines are available to download as Docker containers on Docker Hub (Table 2). On the AWS cloud, data processing pipelines employ Tibanna[50], developed by the 4DN DCIC. Tibanna queries metadata from the 4DN portal to obtain parameters for running the pipelines in the cloud, and updates metadata on the portal upon completion of pipelines to provide access to the processed results, which are stored in Amazon S3. For some of the other assay types, automated cloud pipelines are not yet available; in these cases, data submitters have the option to submit their own processed files that are stored and available in the cloud.

Microscopy Datasets

Given the exciting developments in imaging technologies and their importance in investigating chromosomal dynamics, a considerable part of 4DN efforts are in development and application of such techniques. Accordingly, the 4DN Data Portal was built to also handle microscopy datasets. Complementary to sequencing methods that measure average contact frequencies over large cell populations, imaging-based techniques such as Fluorescence

Table 2. 4D Nucleome Analysis Pipelines.

Pipeline	Steps	Software	Available File Formats
Hi-C[a]	alignment	bwa-mem	.bam
	filtering	pairtools	.pairs
	matrix aggregation and normalization	cooler	.cool, .hic
	merging replicates	cooler	.mcool
MARGI[b]	alignment	bwa-mem	.bam
	filtering	pairtools	.pairs
	merging replicates and matrix aggregation	cooler	.mcool
Repli-seq[c]	alignment	bwa-mem	.bam
	filtering	samtools	-
	binning and aggregation	bedtools	.bw, .bg
CUT&RUN[d]	alignment	bowtie2	.bam
	filtering	Picard, samtools, bedtools	.bg
	peak calling	SEACR	.bw, .bed
ATAC-seq[e]	alignment and filtering	bowtie2	.bed
	peak calling	MACS2	.bw, .bigbed
ChIP-seq[f]	alignment and filtering	bwa	.bed
	peak calling	MACS2, SPP	.bw, .bigbed
RNA-seq[g]	alignment	STAR	.bam
	expression quantification	RSEM	.tsv
	read coverage	STAR	.bw

Listed below are (i) subdirectories for Docker images from https://hub.docker.com/r/4dndcic and (ii) subdirectories for more information from https://data.4dnucleome.org/resources/data-analysis/.
[a] 4dn-hic docker, hi_c-processing-pipeline information.
[b] imargi docker, imargi-pipeline information.
[c] https://hub.docker.com/r/4dndcic/4dn-repliseq, repli-seq-processing-pipeline information.
[d] cut-and-run-pipeline docker, cut-and-run-pipeline information.
[e] encode-atacseq docker, atacseq-processing-pipeline information.
[f] encode-chipseq docker, chipseq-processing-pipeline information.
[g] encode-rnaseq docker, rnaseq-processing-pipeline information.

in situ Hybridization (FISH) allow direct measurement of 3D distances between DNA loci in single cells. In addition their usefulness in cross-validating findings obtained with high-throughput sequencing, microscopy data also provide valuable insight into cell-to-cell variability and dynamics that are lost in experiments with bulk samples.

The 4DN data portal hosts several types of microscopy experiments. To mention a few examples, ChromEMT combines an evolution of electron microscopy tomography with a novel DNA-labeling method, allowing visualization of chromatin organization in interphase and mitotic cells *in situ* at unprecedented resolution[51]. The OptoDroplet assay[52] can assess condensation potential of proteins known to interact with membrane-less organelles in the nucleus[53]. Multiplexed FISH was used to obtain localization data of 8 Mb of human chromosome 19[7]. Single Particle Tracking (SPT) experiments to study nuclear protein dynamics[54] are also present on the portal. High-throughput FISH (hiFISH) represents one example of integrating imaging and sequencing. This technique allows the examination of heterogeneity in genome organization by systematically determining the spatial position and distances between combinations of genomic interaction pairs identified by Hi-C[55].

Together, the 4DN data portal hosts more than 570 microscopy experiment sets at the time of writ-

ing, divided into 25 datasets; a table is available at https://data.4dnucleome.org/microscopy-data-overview. Most of the data are provided as processed files: these are the most reusable and data-rich files and include, for example, locus pair distances from FISH experiments, localization files from high-throughput Small Molecule Switching Nanoscopy[56], or spatio-temporal trajectory coordinates from Single Particle Tracking assays. A selection of raw image files are also available on the 4DN data portal and can be explored directly or downloaded for independent analysis. Microscopy files, unlike sequencing data, do not undergo any automated processing or quality control measures on the data portal at this time.

Biosamples and Tiered Cell Lines

Data hosted at the 4DN data portal come from a variety of biological sources. These sources are mostly human and mouse, but fly, zebrafish, chicken, hamster and green monkey experiments can also be found. The majority of experiments were performed on cell lines, but a few are derived from tissue samples or whole organism samples.

Cell lines at the 4DN data portal can be categorized into three "tiers": Tier 1, Tier 2, and untiered. Early in the consortium, 4DN Network members had decided on a group of cell lines on which to perform coordinated experiments. The goal of this effort was to deliver data that are more directly comparable across different assays, whenever the biological question being studied would allow it. When a biosample is designated as Tier 1 in the 4DN data portal, it means that it is derived from an aliquot of cells obtained from a single common provider, to minimize sample-to-sample variation. Five human cell lines are designated as Tier 1: H1-ESC, GM12878, IMR90, HFF-hTERT (clone 6), and WTC-11. 11 different human and mouse cell lines are currently designed as Tier 2, indicating that multiple 4DN Network investigators have agreed to generate data on them. The untiered cell lines include any cell lines outside of the 4DN-designated lines. A complete list of the 4DN tiered cell lines can be found at https://4dnucleome.org/cell-lines.html.

Data Portal Services

Here, we describe some of the major functionalities of the Data Portal. A key aspect of a useful data portal is to give an overview of the type and amount of data available and allow the user to quickly find the datasets of interest. In addition to uniformly processed raw data, extensive metadata are collected and curated, enabling searches on various aspects of the data.

Browsing Datasets

At the top of the 4DN data portal homepage (https://data.4dnucleome.org/) is a button called "Data" with the following options: Browse All, Browse Sequencing, View Microscopy, and Browse by Publication. Clicking on "Browse All" will load a page of all public Experiment Sets. On the left side of the page is a number of filtering options, where experiments can be filtered on various details. In the example shown (Figure 1), selecting filters of "in situ Hi-C" under experiment type and "Tier 1" under Sample Tier filters results down to a smaller number. Those labeled "Dataset" and "Condition" in the result table are useful in finding datasets of interest. Experiment Sets with the same "Dataset" name generally were generated together as part of a single study or analysis; the "Condition" field explains the differences between them. This is particularly useful when one set of replicates represents a control while another represents a treatment. Together the "dataset" and "condition" fields allow users to easily see which experiment sets go together, and what the experimental differences are.

Each result in the table can be expanded with the '+' button. The expanded view provides access to more details such as available files and the data-generating lab. Files from multiple experiment sets can be downloaded in bulk in the browse view here. Downloads require an account and a corresponding access key, but accounts can be created immediately and are free and accessible to anyone.

Item Pages: Genomics

Clicking on one of the accessions found in the browse results table will bring the user to that experiment set's item page (example shown in Figure 2). More extensive metadata are available on these pages. A reference publication will be featured at the top if the dataset has been published. If the dataset has not been published, it is still available for use, with the following data usage guidelines: (1) the data generating lab should be contacted to discuss possible coordinated publication; (2) the 4DN white paper[1] should be cited; and (3) the lab which generated the data should be acknowledged. Farther below on the page is a pane with several tabs, each providing more details about a different aspect of the experiment set. Typically these include Processed Files, Raw Files, Provenance, Attributions, Details, and Warnings or Commendations. The Raw Files and Processed Files tabs provide metadata about files associated with the experiment set, and also designate the replicate from which they were generated. For processed files this

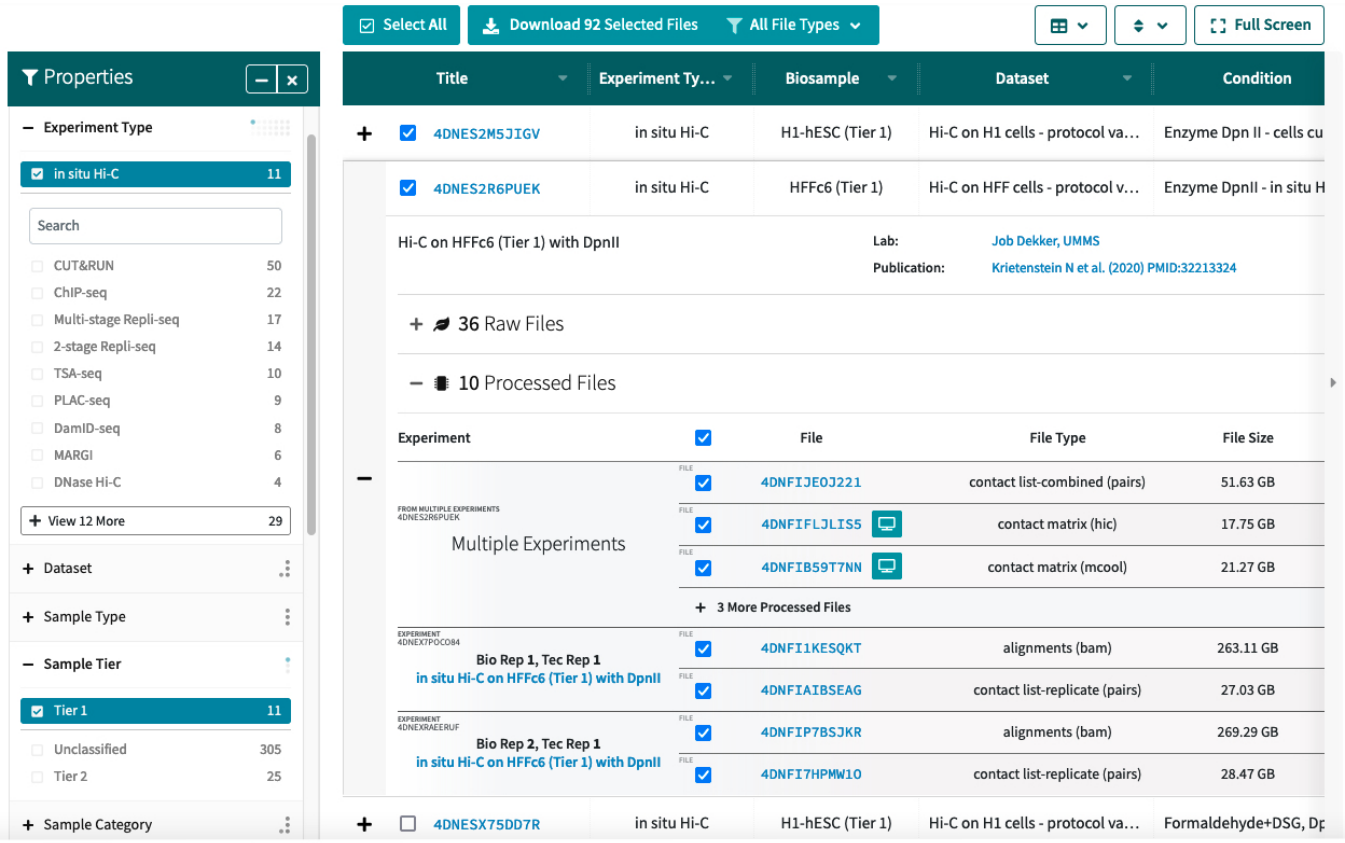

Figure 1. Browsing 4DN Experiment Sets. The browse view features a table of experiment sets, the second of which can be seen expanded here to show additional metadata and information about files. On the left are a number of properties that can be used to filter the results; here "in situ Hi-C" is selected as well as Tier 1 samples. The top two experiment sets in the table are also shown with their checkboxes checked, so that the "Download Files" button above the table can be used.

tab often includes a small HiGlass view of the final processed output, which can be expanded into a new window. If processed files were generated by one of the 4DN processing pipelines, there will also be a provenance graph available to view in a subsequent tab. This directed graph shows the inputs and outputs for each step in the pipeline that was run on the experiment set. In the case of user-submitted processed files where a 4DN processing pipeline was not run, a provenance graph will still be present, but it will show only relationships between input and output files.

The final tab on item pages for Experiment Sets, Experiments, and Biosamples will be Warnings or Commendations. Presently these are used to indicate if an experiment lacks replicates, or to indicate whether the biosamples have all of the requisite metadata. The biosamples with a gold commendation have extensive metadata information as well as a morphology image taken before harvesting. This allows people who want to reuse the data to know more about the cultures and whether they had been growing with typical morphology. If a sample doesn't meet one or more metadata requirements for the gold commendation, this will show up under Warnings with a message indicating which piece(s) of metadata is missing.

All item types will also have Attributions and Details tabs. The attribution tab indicates which lab(s) generated the data and any associated publications. Datasets published in journals often must also be deposited in GEO or SRA, and some 4DN datasets have also been deposited to the ENCODE portal; if available, identifiers to other databases that may also host the data are also included here. The details tab has a list of all metadata fields associated with the item, regardless of whether or not it is already displayed elsewhere on the page.

Comparing Files in HiGlass

Whereas several genomic and epigenomic data repositories have a 1D genome browser tool, contact frequencies across distant genomic locations require 2D visualization. To this end, a HiGlass browser[8] has been integrated into

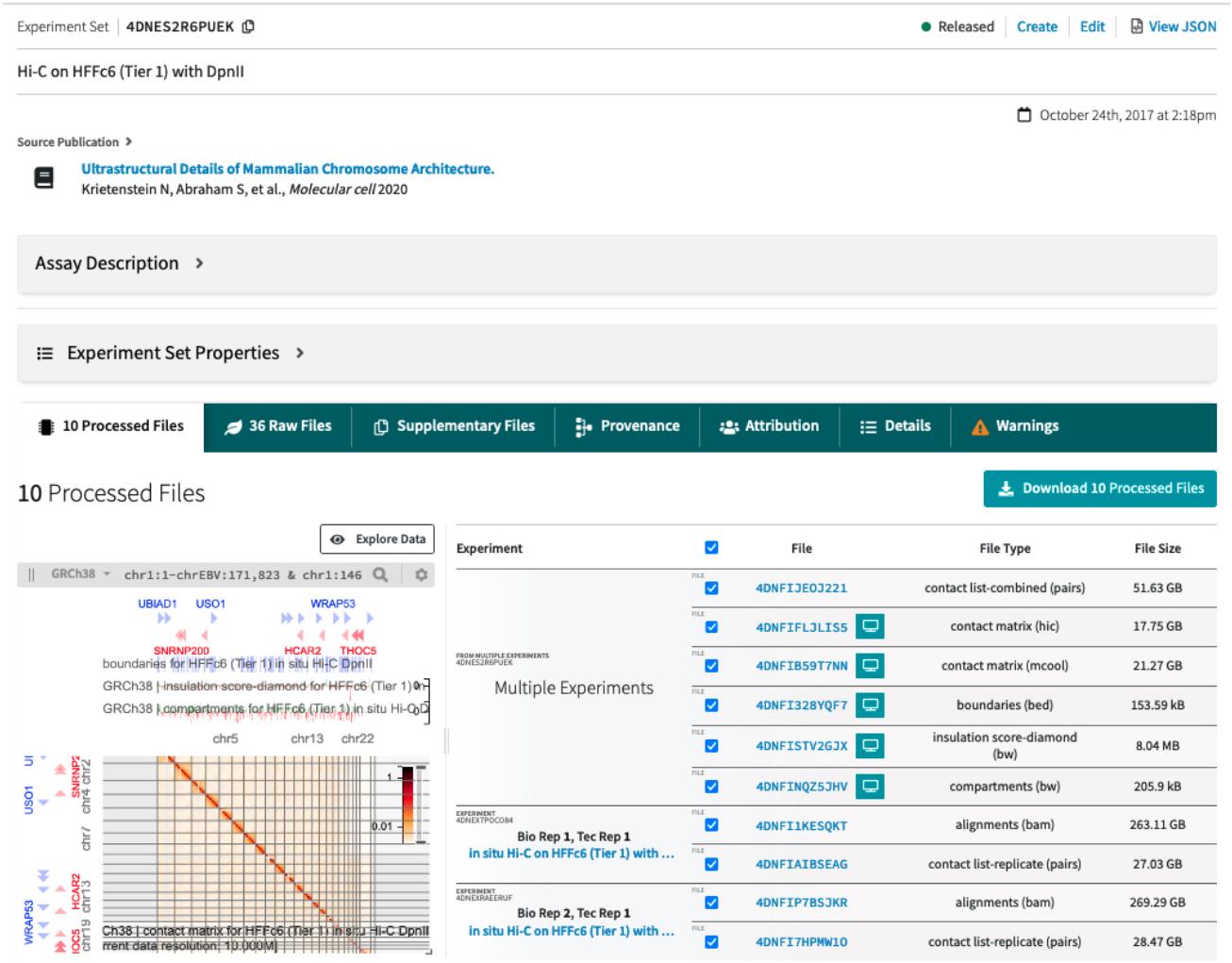

Figure 2. Item Page for a Replicate Set of Hi-C experiments. A source publication is shown near the top of the page, when available. Below this are two dropdown boxes: the first, titled Assay Description, can be expanded for an explanation of the assay, and the second, titled Experiment Set Properties, contains a selection of basic metadata fields. Below these is a window with several tabs. The selected tab shows the processed files associated with the experiment. The .mcool contact matrix file is visualized on the left using the integrated HiGlass browser as a 2D track, with TAD boundaries, insulation scores, and compartments as 1D tracks above it. This display can be expanded for further data exploration. Scrolling down on this page would reveal the quality control metrics associated with the processed data.

the portal for interactive visualization of both 2D contact matrices and 1D genomic tracks. The 4DN visualization workspace can be accessed either at https://data.4dnucleome.org/tools/visualization or by clicking on the "Explore Data" button above the miniature HiGlass view displayed on an Experiment Set item page.

The close integration of HiGlass into the portal allows users to leverage the portal search capabilities to identify files of interest for visualization that can then be added to existing HiGlass views. Files can be added to the HiGlass display by clicking on the Add Data button, which opens a window where users can filter via various metadata fields to find a file of interest. Multiple 2D contact matrices can be compared at once (Figure 3), and the position and zoom level of each view is locked to the first one, such that zooming in or out or dragging the mouse to different genomic locations moves all the matrices in unison.

1D tracks can also be added to each HiGlass view. By default, each 1D track added will be added to every view in the display. This allows it to be compared to each 2D matrix simultaneously. Multiple 1D tracks can be added in this fashion. Users can then save their customized views for later retrieval and sharing with collaborators.

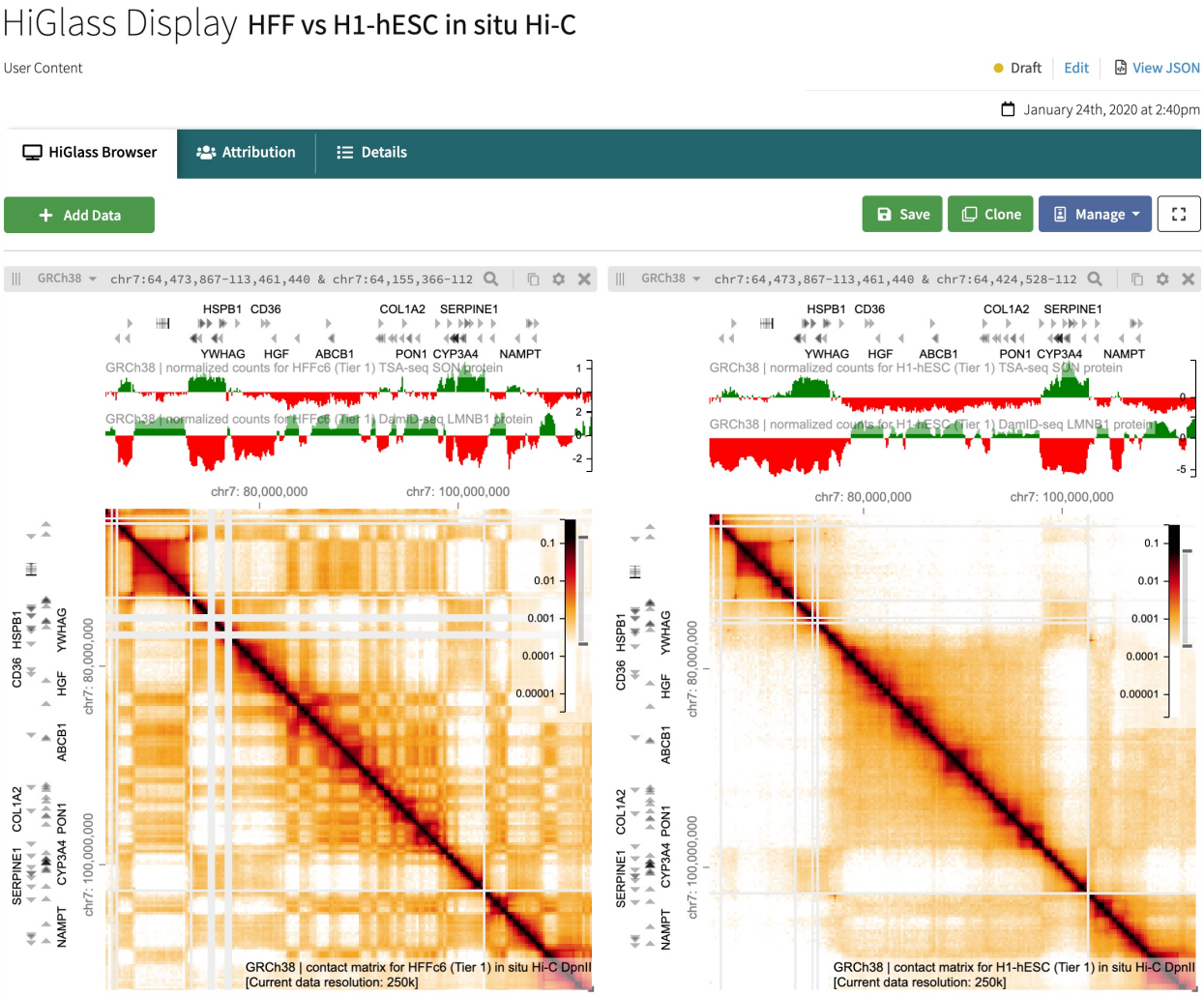

Figure 3. HiGlass Display containing 2D and 1D Genomic Tracks. The display shows visualization of a HFF in situ Hi-C contact matrix on the left, and one for H1 hESCs on the right. Above each 2D contact matrix visualization are 1D tracks from a TSA-seq experiment (topmost) and a DamID-seq experiment (second from top), in the cell type that matches the corresponding matrix. The "Add Data" button in the top left gives the user the option to add more files to the visualization display; on the top right, the user has the ability to save the display, clone the display to create a new item, or to manage permissions of who can view the display.

Finding Microscopy Experiments

In the Data menu in the top navigation bar, clicking on "View Microscopy Datasets" will bring the user to a table of the microscopy datasets that can be found in the portal. These are organized by Dataset label rather than by Experiment Set, as some datasets comprise large numbers of replicate sets, allowing the table to be much more concise and straightforward. The second column contains a link to a browse page of all the Experiment Sets in the dataset.

A microscopy experiment set item page (Figure 4) is organized much the same as the sequencing experiment sets, with a few notable differences. In the overview section above the tabs, in addition to common metadata there may be 2 extra fields: a sample image, for sets that include raw files, and information about the imaging paths. The Imaging Path refers to metadata about what was imaged, including cellular target, antibodies or probes, and the detected label. In addition, there are no processing pipelines run by the 4DN data portal on microscopy datasets, so the Provenance tab will often be dimmed indicating no provenance is available. However, in some cases submitting labs indicate which files are processed results derived from others and in these cases a simple provenance graph reflecting this is shown.

If a sample image is present and derives from a raw image file on the portal rather than a rendered jpg, then

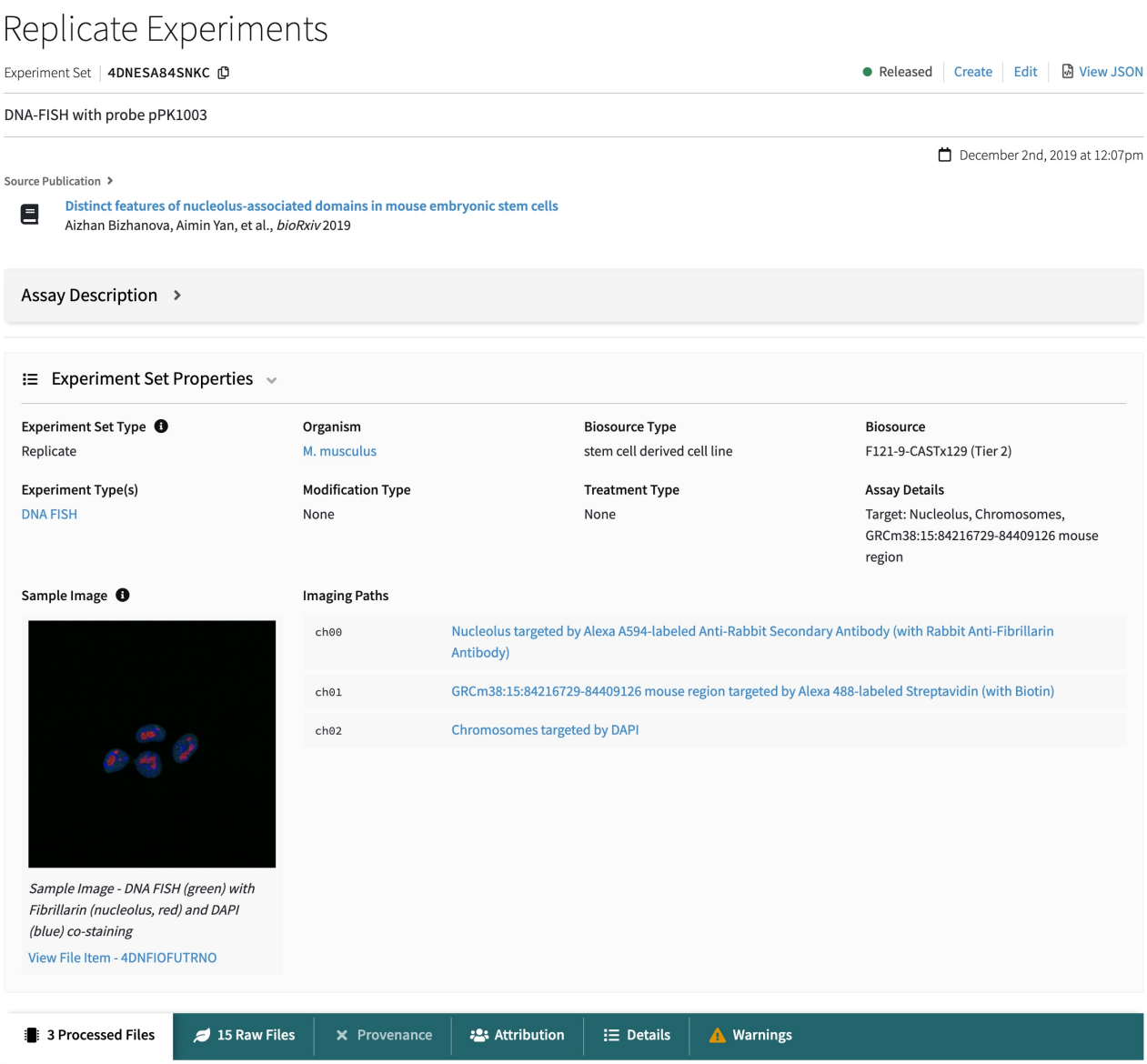

Figure 4. Microscopy Experiment Set item page. Item pages for microscopy experiment sets are similar to those for genomics, but with a few important differences. The sample image field contains an image preview, which can be clicked for a popup containing an interactive image viewer. Beside the sample image is a field called "Imaging Paths", which describes what is imaged in each channel for the experiments, including the biological target and the antibodies or probes used.

clicking on the image will bring the user to an interactive image viewer, where it is possible to scroll through focal planes in a z-stack, or through time points in a time-lapse image, or adjust signal and color levels in the display.

Portal Architecture

The 4DN Consortium was projected to generate large volumes of diverse datasets at its initiation. To be useful to scientists with different technical and scientific backgrounds, primary objectives were to ensure users can determine what data are available on the portal, and that data of different modalities should be made accessible to all scientists. This requires an architecture that is modular, responsive, and scalable, with an easily extensible data model. Our software architecture is entirely cloud-based, as we envisioned that taking analytical tools to the cloud environment where data reside, rather than downloading large amounts of data to a local server, is quickly becoming the preferred mode of data analysis.

The design of the 4DN data portal infrastructure (Figure 5), originally based on the ENCODE infrastructure, includes the following components: (1) A postgres database storing metadata in json format, first developed in ENCODE; (2) The python pyramid framework for the database known as SnoVault[57], first developed in ENCODE but further tailored and developed by the 4DN DCIC (https://github.com/4dn-dcic/snovault); (3) The FourFront front-end, originally based on EncodeD[57] from ENCODE, but engineered by 4DN DCIC to feature a data model for representing diverse datasets, and includes a modern front-end with reactJS to provide a responsive user experience; (4) Elasticsearch that provides fast and efficient search with various metadata fields by indexing all items and formatting them for retrieval; (5) AWS S3 used for file storage, enabling all public data files to be accessed via the data portal interface; (6) A RESTful API underlying the infrastructure, through which all metadata in the portal can be accessed.

In the AWS cloud environment, changes in the database are indexed within seconds to minutes, allowing updates and releases of datasets with minimal overhead. Similarly, the infrastructure was developed with features that allow the release of changes to data model and software versions overnight without any server downtime. Thus, we are able to ingest and release datasets with new data models any time, without having to enforce data release cycles.

All of the software used in the data portal is open access and open source (https://github.com/4dn-dcic/). Finally, the 4DN data portal also has an associated support contact email (support@4dnucleome.org), allowing curators to quickly address any question or concern from end users.

Discussion

With 1226 sequencing and 576 microscopy experiment sets that are publicly available and consistently curated, the portal provides a high volume of data in a wide array of data types to nuclear architecture researchers across the globe. The portal interface provides user-friendly search capabilities with many options for filtering results, and the result pages have been designed to communicate extensive metadata for each item. Uniform processing or assays with automated pipelines also ensures that the processed results are consistent and directly comparable.

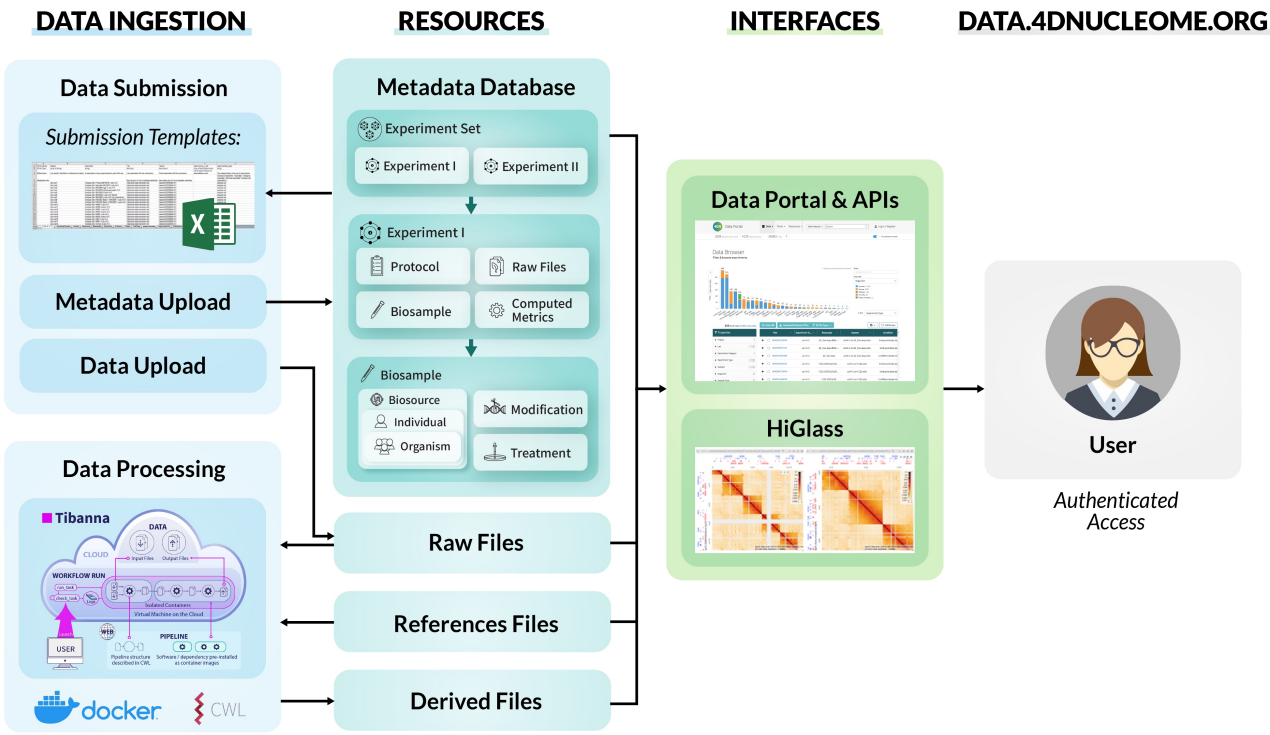

Figure 5. 4DN Data Portal Architecture. Metadata are submitted by users via spreadsheet forms and gets loaded into the database, while associated files are uploaded into cloud storage. Once metadata and files are in place, automated processing pipelines can be run on AWS using the Tibanna pipeline runner. All metadata can then be searched via the data portal website, where files can also be visualized with HiGlass. The data portal website is accessible to external users, who can also log in in order to download files.

The integrated HiGlass browser allows the data portal to serve not only as a hub for searching and downloading research data but also as a visualization platform. 2D output from genomic assays like Hi-C can be directly compared to other 2D assays like DNA SPRITE, or to results from 1D assays like ChIP-seq or CUT&RUN. Even when only looking at Hi-C experiments, some 4DN data contributors have uploaded experiments comparing different variations in Hi-C protocols[11], and the HiGlass browser provides an easy way to examine differences in data output resulting from protocol modifications.

Prior to the development of the 4DN data portal, there was a lack of a centralized repository that specializes in chromatin conformation and nuclear structure data, specifically. Other repositories that host Hi-C and similar datasets include the ENCODE portal[58], as well as NCBI's GEO[59] and EMBL-EBI's ArrayExpress[60]. GEO and ArrayExpress are great resources for hosting published sequencing datasets, but they lack a finer specialized focus or visualization capabilities for processed results. Their metadata also tends to be minimal and less structured, which is understandable given the huge diversity of datasets that need to be hosted there. ENCODE has an infrastructure that is very similar to the 4DN data portal, and it has amassed an impressive number of sequencing datasets over the years it has been active. Although they have Hi-C datasets available, since their portal has been engineered to focus more on 1D datasets, they lack a 2D genome browser and the Hi-C datasets cannot be visualized there. The 4DN data portal thus fills a very important niche: it represents a centralized repository for nuclear structure data specifically, with a data model tailored to provide extensive metadata about each experiment; it enables visualization of processed results, where 2D genomic data can be compared to 1D genomic tracks; and it also houses microscopy data that can complement sequencing results. To our knowledge this is unique to the 4DN data portal.

The second phase of the NIH Common Fund 4D Nucleome Project began in 2020, and will proceed for five years, with a focus on the role of nuclear structure and function in human health and disease, as well as on the continued development of data visualization and integration tools. Thus, we expect the data volume in the 4DN data portal to continue to increase. Development on the portal is ongoing to ensure that the 4DN data portal continues to serve the needs of the nuclear biology research community.

Data Availability

All datasets described are available at https://data.4dnucleome.org/.

Code Availability

All of the software that comprises the 4DN data portal infrastructure is free and open source. All code repositories mentioned are available from https://github.com/4dn-dcic.

Acknowledgments

We thank J. Michael Cherry and the development team at ENCODE for helpful discussions and advice regarding setting up and running a data portal, as well as allowing us to build off of the ENCODE software ecosystem. This work was funded by the NIH Common Fund grant 1U01CA200059 to P.J.P.

Authors' Contributions

S.B.R. wrote the manuscript with the help of A.J.S., A.C., K.K., C.B., L.M., B.H.A. and P.J.P. J.J., A.K.B., C.V., S.B.R., A.J.S., K.K., A.C., C.B., L.M., S.L., A.D.V., W.R., K.M.P., P.K., N.A., M.I., S.U.O. and U.C. developed the software. A.J.S., K.K., A.C., S.B.R., and L.M. handled data submissions. S.R.E. carried out graphic design. P.J.P. supervised the work along with N.G., L.A.M., and B.H.A.

Competing Interests

The authors declare no conflict of interest.

References

1. Dekker, J. *et al.* The 4D Nucleome Project. *Nature* **549,** 219–226 (2017).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
3. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 139–144 (2010).
4. Van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nat Biotechnol* **18,** 424–428 (2000).
5. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6.** doi:10.7554/eLife.21856 (2017).

6. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174,** 744–757.e24 (2018).
7. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet* **14,** e1007872 (2018).
8. Kerpedjiev, P. *et al.* HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol* **19.** doi:10.1186/s13059-018-1486-1 (2018).
9. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162,** 108–119 (2015).
10. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol* **16,** 152 (2015).
11. Akgol Oksuz, B. *et al.* Systematic evaluation of chromosome conformation capture assays. *Nat Methods* **18,** 1046–1055 (2021).
12. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502,** 59–64 (2013).
13. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat Methods* **14,** 263–266 (2017).
14. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544,** 110–114 (2017).
15. Fullwood, M. J. *et al.* An Oestrogen Receptor alpha-bound Human Chromatin Interactome. *Nature* **462,** 58–64 (2009).
16. Li, X. *et al.* Long-read ChIA-PET for base-pair resolution mapping of haplotype-specific chromatin interactions. *Nat Protoc* **12,** 899–915 (2017).
17. Zheng, M. *et al.* Multiplex chromatin interactions with single-molecule precision. *Nature* **566,** 558–562 (2019).
18. Fang, R. *et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26,** 1345–1348 (2016).
19. Sridhar, B. *et al.* Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Curr Biol* **27,** 602–609 (2017).
20. Yan, Z. *et al.* Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proc Natl Acad Sci USA* **116,** 3328–3337 (2019).
21. Wu, W. *et al.* Mapping RNA–chromatin interactions by sequencing with iMARGI. *Nat Protoc* **14,** 3243–3272 (2019).
22. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* **217,** 4025–4048 (2018).
23. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.* arXiv: 1303.3997 (2013).
24. Goloborodko, A., Abdennur, N., Venev, S., Hbbrandao & Gfudenberg. *mirnylab/pairtools v0.3.0* version v0.3.0. 2019. doi:10.5281/ZENODO.2649383. https://zenodo.org/record/2649383 (2019).
25. Andrews, S. *FastQC: a quality control tool for high throughput sequence data.* 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
26. Lee, S., Vitzthum, C., Alver, B. H. & Park, P. J. *Pairs and Pairix: a file format and a tool for efficient storage and retrieval for Hi-C read pairs* preprint (2021). doi:10.1101/2021.08.24.457552.
27. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36,** 311–316 (2020).
28. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47,** 598–606 (2015).
29. Marchal, C. *et al.* Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nat Protoc* **13,** 819–839 (2018).
30. Akhtar, W. *et al.* Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* **154,** 914–927 (2013).
31. Vertii, A. *et al.* Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. *Genome Res* **29,** 1235–1249 (2019).
32. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543,** 519–524 (2017).
33. Belaghzal, H. *et al.* Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. *Nat Genet* **53,** 367–378 (2021).
34. Paulsen, M. T. *et al.* Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **67,** 45–54 (2014).
35. Tavares-Cadete, F., Norouzi, D., Dekker, B., Liu, Y. & Dekker, J. Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nat Struct Mol Biol* **27,** 1105–1114 (2020).

36. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3,** 99–101 (2016).

37. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485,** 376–380 (2012).

38. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* **485,** 381–385 (2012).

39. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nature Genetics* **52,** 8–16 (2020).

40. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Science Advances* **5,** eaaw1668 (2019).

41. Zufferey, M., Tavernari, D., Oricchio, E. & Ciriello, G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology* **19,** 217 (2018).

42. Venev, S. *et al. open2c/cooltools: v0.4.1* version v0.4.1. 2021. doi:10.5281/ZENODO.5214125. https://zenodo.org/record/5214125.

43. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523,** 240–244 (2015).

44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

45. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).

47. Institute, B. *Picard toolkit* version 2.20.7. 2019. https://broadinstitute.github.io/picard/.

48. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics & Chromatin* **12,** 42 (2019).

49. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46,** D794–D801 (Jan. 4, 2018).

50. Lee, S. *et al.* Tibanna: software for scalable execution of portable pipelines on the cloud. *Bioinformatics* **35,** 4424–4426 (2019).

51. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357.** doi:10.1126/science.aag0025 (2017).

52. Shin, Y. *et al.* Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets. *Cell* **168,** 159–171.e14 (2017).

53. Courchaine, E. M. *et al.* DMA-tudor interaction modules control the specificity of in vivo condensates. *Cell* **184,** 3612–3625.e17 (2021).

54. Hansen, A. S. *et al.* Robust model-based analysis of single-particle tracking experiments with Spot-On. *Elife* **7.** doi:10.7554/eLife.33125 (2018).

55. Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* **176,** 1502–1515.e10 (2019).

56. Barentine, A. E. S. *et al.* 3D Multicolor Nanoscopy at 10,000 Cells a Day. *bioRxiv,* 606954 (2019).

57. Hitz, B. C. *et al.* SnoVault and encodeD: A novel object-based storage system and applications to ENCODE metadata. *PLoS One* **12.** doi:10.1371/journal.pone.0175310 (2017).

58. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44,** D726–D732 (2016).

59. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* **41,** D991–995 (2013).

60. Parkinson, H. *et al.* ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* **39,** D1002–1004 (2011).