Genomic resources

Mikhail Dozmorov

Spring 2018

High-throughput data repositories

- **GEO**: Gene Expression Omnibus.
  - Host array- and sequencing-based data.
- **ArrayExpress**: European version of GEO.
  - Better curated than GEO but has less data.
- **SRA**: Sequence Read Archive.
  - Designed for hosting large scale high-throughput sequencing data, e.g., high speed file transfer.
  - Data are required to be deposited in one of the databases when paper is accepted

Sequence Read Archive (SRA)

- The NCBI database which stores sequence data obtained from next generation sequence (NGS) technology
  - Archives raw NGS data for various organisms from several platforms (FASTQ files)
  - Serves as a starting point for "secondary analyses"
  - Provides access to data from human clinical samples to authorized users who agree to the datasets' privacy and usage mandates
- Search metadata to locate the sequence reads for download and further downstream analyses

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?

https://www.ncbi.nlm.nih.gov/sra/

Getting data from SRA

- The NCBI `sratoolkit` provides two command line tools to allow local BLAST searches against specific sra files directly
- `fastq-dump`: Convert SRA data into fastq format
- `prefetch`: Allows command-line downloading of SRA, dbGaP, and ADSP data
- `sam-dump`: Convert SRA data to sam format
- `sra-pileup`: Generate pileup statistics on aligned SRA data
- `vdb-config`: Display and modify VDB configuration information
- `vdb-decrypt`: Decrypt non-SRA dbGaP data ("phenotype data")

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

Getting data from SRA

- `.sra` files are NOT FASTQ files - need to further convert them using `sratoolkit`

```
wget ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP101/SRP101962/SRR5346141/SRR5346
# To split paired-end reads, use -I option
sratoolkit.2.8.1-win64/bin/fastq-dump -I --split-files SRR5346141
```

https://www.ncbi.nlm.nih.gov/books/NBK47528/

Long reads

- Bacterial and eukaryotic genomes available from PacBio DevNet

https://github.com/PacificBiosciences/DevNet/wiki/Datasets

Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J, et al. 2014. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data 1: 140045.

UCSC Genome Browser

- The UCSC genome browser is a graphical viewer for visualizing genome annotations.
- Initially developed by Jim Kent on 2000 when he was a Ph.D. student in Biology.
- Host genomic annotation data for many species.
- Provide other tools for genomic data analysis and interfaces for querying the database.

http://genome.ucsc.edu/

UCSC Genome Browser Track Hubs

- Track hubs are web-accessible (HTTP or FTP) directories of genomic data that can be viewed on the UCSC Genome Browser
- Tracks can be aggregated using a text document in the UCSC Genome Browser track hub format
  - Advantage: Can be easily distributed to collaborators / users of your resources
  - Disadvantage: Need to generate this text document

http://genome.ucsc.edu/goldenpath/help/hgTrackHubHelp.html

Small track hub example

- Minimum set of track description fields:
  - *track* - Symbolic name of the track
  - *type* - One of the supported formats
    - bigWig, bigBed, bigGenePred, bam, vcfTabix ...
  - *bigDataUrl* - Web location (URL) of the data file
  - *shortLabel* - Short track description (Max 17 characters)
  - *longLabel* - Longer track description (displayed over tracks in the browser)

Small track hub example

```
track McGill_MS000101_monocyte_RNASeq_signal_forward
type bigWig
bigDataUrl http://epigenomesportal.ca/public_data/MS000101.monocyte.RNASeq.signal_forward.bigWig
shortLabel 000101mono.rna
longLabel MS000101 | human | monocyte | RNA-Seq | signal_forward

track McGill_MS000101_monocyte_RNASeq_signal_reverse
type bigWig
bigDataUrl http://epigenomesportal.ca/public_data/MS000101.monocyte.RNASeq.signal_reverse.bigWig
shortLabel 000101mono.rna
longLabel MS000101 | human | monocyte | RNA-Seq | signal_reverse
```

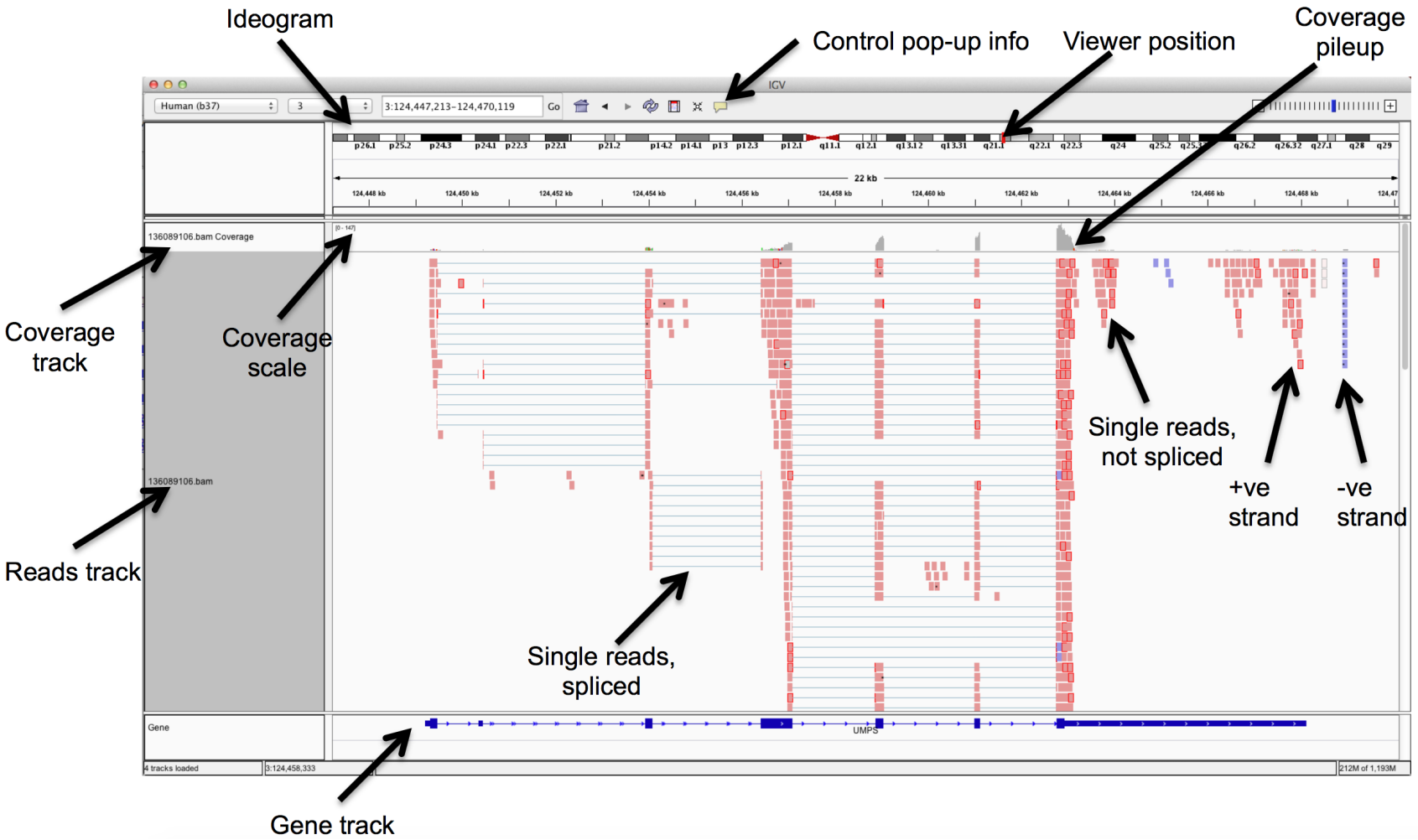
WashU Epigenome Browser

- Visualizing (Epi)Genomics Data
- Includes Roadmap Epigenome data
- Supports many track types included in the UCSC Browser
- Can also load UCSC track hub documents

https://epigenomegateway.wustl.edu/

Visualization

Integrative Genomics Viewer (IGV)



http://software.broadinstitute.org/software/igv/

IGV

Features

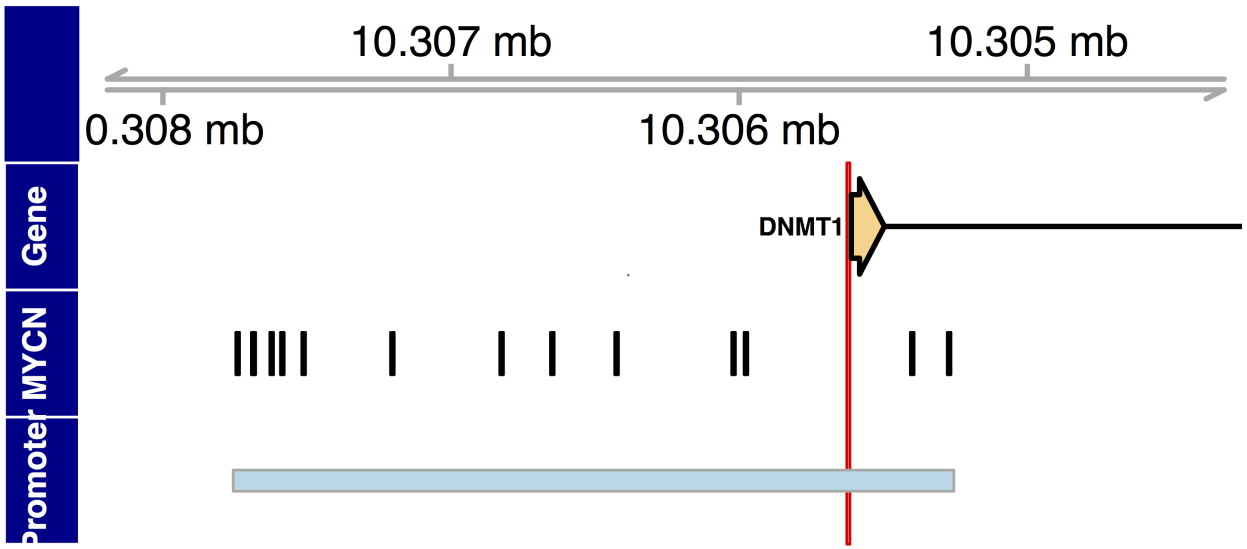
- Explore large genomic datasets with an intuitive, easy-to-use interface.
- Integrate multiple data types with clinical and other sample information.
- View data from multiple sources:
  - local, remote, and "cloud-based".
  - Intelligent remote file handling - no need to download the whole dataset
- Automation of specific tasks using command-line interface

Tutorial: https://github.com/griffithlab/rnaseq_tutorial/wiki/IGV-Tutorial

Gviz R package

- Plotting data and annotation information along genomic coordinates
- Track-oriented

Figure. Promoter of DNMT1 (+500bp .. -2000bp around transcription start site (TSS), blue bar) contains numerous MYCN canonical binding sites (JASPAR ID MA0104.4, black ticks). Vertical red line marks the TSS of DNMT1.



https://bioconductor.org/packages/release/bioc/html/Gviz.html

epivizR R package

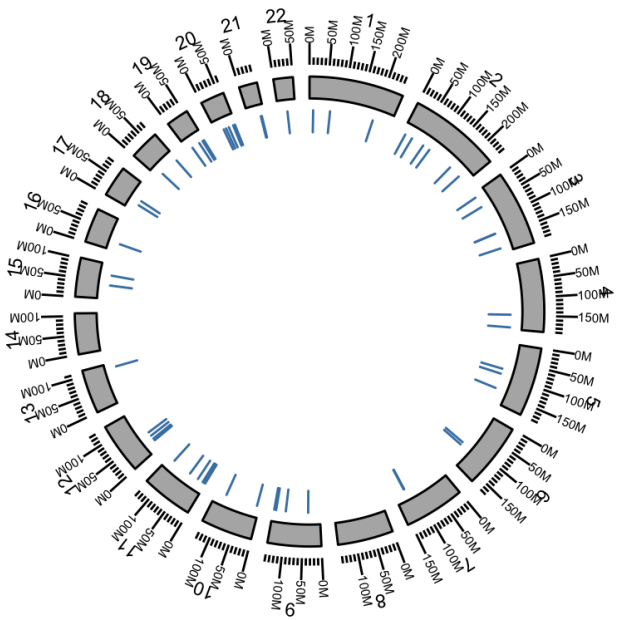
- D3-based interactive visualization tool for functional genomics data.
- Multiple visualizations using scatterplots, heatmaps and other user-supplied visualizations.
- Includes data from the Gene Expression Barcode project for transcriptome visualization.

http://epiviz.cbcb.umd.edu/

https://epiviz.github.io/

ggbio R package

- ggplot2 for genomic data



https://www.bioconductor.org/packages/release/bioc/html/ggbio.html

http://www.sthda.com/english/wiki/ggbio-visualize-genomic-data

Other visualization tools

- Review of omics data visualization tools, summary table: Schroeder, Michael P., Abel Gonzalez-Perez, and Nuria Lopez-Bigas. "Visualizing Multidimensional Cancer Genomics Data." Genome Medicine 5, no. 1 (2013): 9. https://doi.org/10.1186/gm413.
- GIVE (Genomic Interaction Visualization Engine) - an open source programming library that allows anyone with HTML programming experience to build custom genome browser websites or apps

https://genomemedicine.biomedcentral.com/articles/10.1186/gm413

Cao, Xiaoyi, Zhangming Yan, Qiuyang Wu, Alvin Zheng, and Sheng Zhong. "Building a Genome Browser with GIVE." BioRxiv, January 1, 2018. https://doi.org/10.1101/177832. https://zhong-lab-ucsd.github.io/GIVE_homepage/

Other genome browsers/databases

General

- NCBI Genome Data Viewer, https://www.ncbi.nlm.nih.gov/genome/gdv/
- Ensembl genome browser, https://www.ensembl.org/

Species-specific genome browser

- MGI: Mouse genome informatics, http://www.informatics.jax.org/
- **wormbase** http://www.wormbase.org/
- **Flybase** http://flybase.org/
- **SGD** (yeast) https://www.yeastgenome.org/
- **TAIR DB** (arabidopsis) https://www.arabidopsis.org/
- **MBGD microbial genome database** http://mbgd.genome.ad.jp/

High-throughput data repositories

- **TCGA** (The Cancer Genome Atlas) data portal, https://cancergenome.nih.gov/
  - Host data generated by TCGA, a big consortium to study cancer genomics.
  - Huge collection of cancer-related data: different types of genomic, genetic and clinical data for many different types of cancers.
- **ENCODE** (the ENCyclopedia Of DNA Elements) data coordination center (http://genome.ucsc.edu/ENCODE/):
  - Host data generated by ENCODE, a big consortium to study functional elements of human genome.
  - Rich collection of genomic and epigenomic data.

Connectivity Map

- **Connectivity Map** - 4 cell lines and 1309 perturbagens at several concentrations. Gene expression change after treatment
- **Connectivity Map 2** - 1,319,138 L1000 profiles from 42,080 perturbagens

https://portals.broadinstitute.org/cmap/forceLogin.jsp

Subramanian, Aravind, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." Cell 171, no. 6 (November 2017): 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

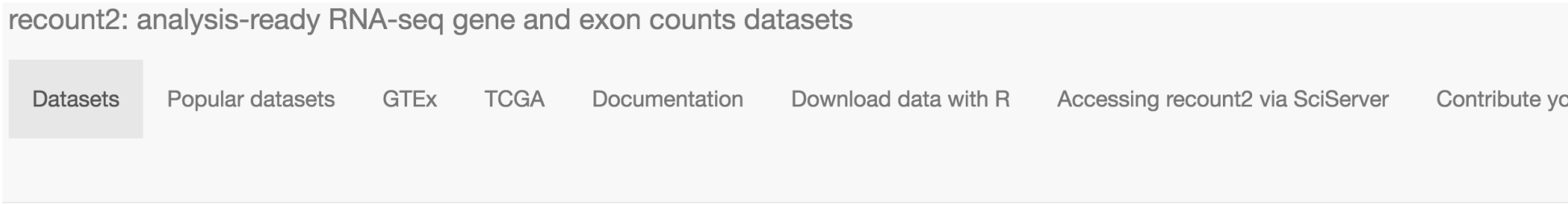
https://clue.io/

API access, https://clue.io/api

Many analytical tools, http://lincsproject.org/

Query your up/downregulated genes, https://clue.io/l1000-query

RECOUNT2 - A multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets



recount2 A multi-experiment resource of analysis-ready RNA-seq gene count datasets

- Uniformly processed (Rail-RNA) gene- and exon counts
- Signal coverage in bigWig format
- Phenotype data
- RangedSummarizedExperiment R objects

Web: https://jhubiostatistics.shinyapps.io/recount/

R package: https://bioconductor.org/packages/release/bioc/html/recount.html

Collado-Torres, Leonardo, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, and Jeffrey T Leek. "Reproducible RNA-Seq Analysis Using Recount2." Nature Biotechnology 35, no. 4

ARCHS4 - all RNA-seq and ChIP-seq sample and signature search

- A web resource that makes the majority of previously published RNA-seq data from human and mouse freely available at the gene count level
- All available FASTQ files from RNA-seq experiments were retrieved from the Gene Expression Omnibus (GEO) and aligned using a cloud-based infrastructure.
- 72,363 mouse and 65,429 human samples
- Gene-centric exploratory analysis of average expression across cell lines and tissues, top co-expressed genes, and predicted biological functions and protein-protein interactions for each gene based on prior knowledge combined with co-expression
- Processed data in HDF5 format

Lachmann, Alexander, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hyojin J. Lee, Moshe C. Silverstein, Lily Wang, and Avi Ma'ayan. "Massive Mining of Publicly Available RNA-Seq Data from Human and Mouse." BioRxiv, January 1, 2017. https://doi.org/10.1101/189092.

http://amp.pharm.mssm.edu/archs4/download.html

ExperimentHub

- ExperimentHub provides a central location where curated data from experiments, publications or training courses can be accessed.
- Each resource has associated metadata, tags and date of modification.
- The R package client creates and manages a local cache of files retrieved enabling quick and reproducible access.
- Usage similar to `AnnotationHub`

https://bioconductor.org/packages/release/bioc/html/ExperimentHub.html

Galaxy

- Web-based framework offering a user-friendly interface mapping to most popular bioinformatics tools
  - "Data intensive biology for everyone."
- Allows for reproducible results
  - Steps / parameters kept in history
- Ability to design custom pipelines and import others'
  - All through a user-friendly GUI
- Tailored for small/medium scale projects with not too many samples

https://usegalaxy.org/

Other resources

- **BaseSpace** - Illumina-oriented cloud computing environment, https://basespace.illumina.com/home/index
- **GenePattern** - web-based computational biology suite of tools for genomic analysis. http://software.broadinstitute.org/cancer/software/genepattern/
- **GenomeSpace** - integrated environment of the aforementioned genomic platforms allowing the data to be stored in one place and analyzed by a multitude of tools. http://www.genomespace.org/

Side-by-side comparison of many resources
https://docs.google.com/spreadsheets/d/1o8iYwYUy0V7IECmu21Und3XALwQihioj23WGv-w0itk/pubhtml

Summarized data sets, services and resources

Name	Website	Notes
ArrayExpress[95]	www.ebi.ac.uk/arrayexpress	Archives processed data from high-throughput functional genomics experiments
Beacon	beacon-network.org	Platform for sharing genetic mutations across web services called 'beacons'
Bravo	bravo.sph.umich.edu	TOPMed data browser for accessing alleles across over 60,000 whole genomes
Expression Atlas[121]	www.ebi.ac.uk/gxa	Gene expression information across 3,000 transcriptomic experiments from ArrayExpress
PCAWG	docs.icgc.org/pcawg	Called germline and somatic variants, including structural variants, from over 5,600 tumour and normal samples across ICGC projects
recount2[61]	jhubiostatistics.shinyapps.io/recount	Web and R/Bioconductor resource for accessing genome coverage data from over 70,000 archived human RNA-seq samples, including publicly available SRA, TCGA and GTEx samples
RNASeq-er[93]	www.ebi.ac.uk/fg/rnaseq/api	Provides programmatic access to processed outputs for all archived publicly available RNA-seq samples
Snaptron	snaptron.cs.jhu.edu	Allows rapid querying of splice junctions, splicing patterns and metadata from recount2
Tatlow-Piccolo[68]	osf.io/gqrz9	Quantified transcripts across TCGA and CCLE
Toil[63]	xenabrowser.net/data-pages/?host=https://toil.xenahubs.net	Processed outputs from over 20,000 RNA-seq samples including TCGA and GTEx
Xena[94]	xena.ucsc.edu	Visualizes investigators' new functional genomics data next to publicly available data

CCLE, Cancer Cell Line Encyclopedia; GTEx, Genotype-Tissue Expression Project; ICGC, International Cancer Genome Consortium; PCAWG, Pan-Cancer Analysis of Whole Genomes; RNA-seq, RNA sequencing; SRA, Sequence Read Archive; TCGA, The Cancer Genome Atlas; TOPMed, Trans-Omics for Precision Medicine.

Langmead, Ben, and Abhinav Nellore. "Cloud Computing for Genomic Data Analysis and Collaboration." Nature Reviews Genetics, January 30, 2018. https://doi.org/10.1038/nrg.2017.113.

www.ebi.ac.uk/arrayexpress, beacon-network.org, bravo.sph.umich.edu, www.ebi.ac.uk/gxa, docs.icgc.org/pcawg, jhubiostatistics.shinyapps.io/recount, www.ebi.ac.uk/fg/rnaseq/api, snaptron.cs.jhu.edu, osf.io/gqrz9, xenabrowser.net/datapages/?host=https://toil.xenahubs.net, xena.ucsc.edu