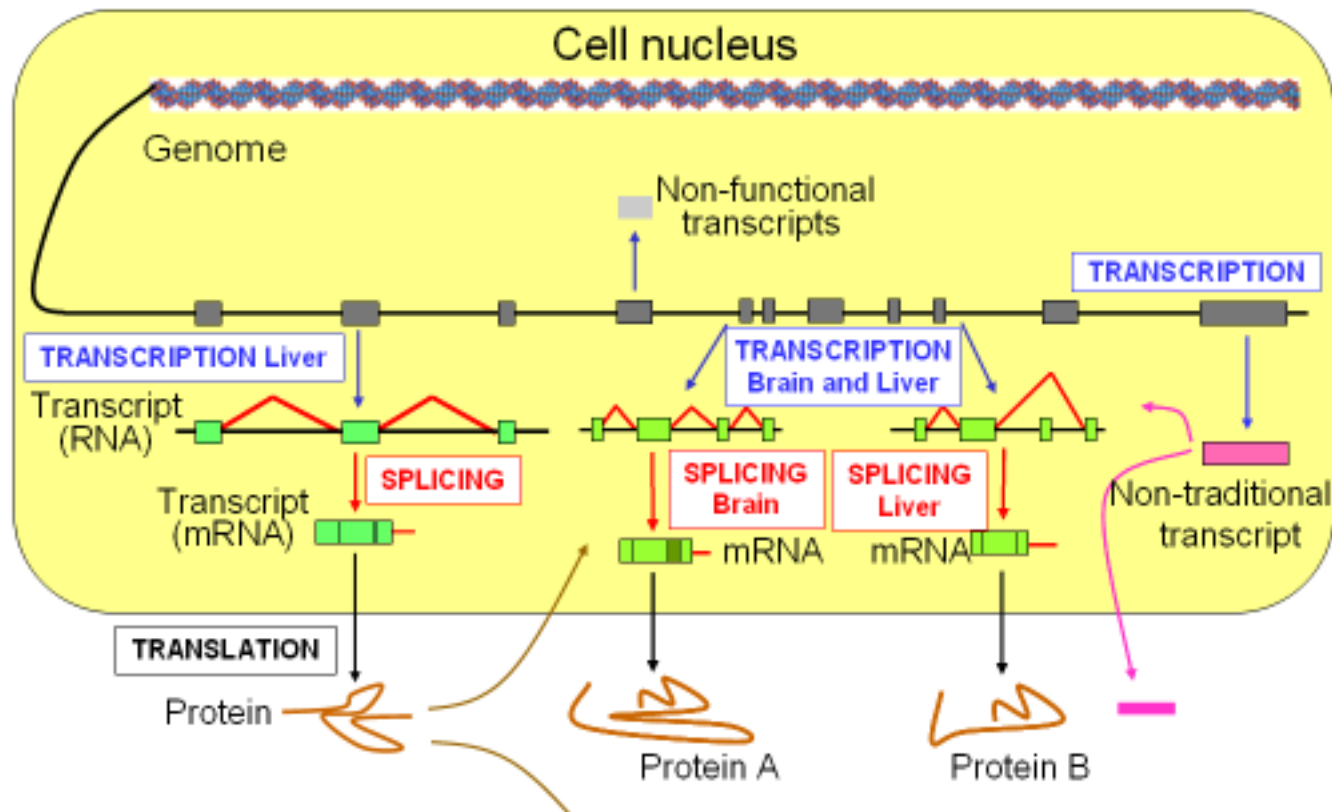


Towards Large-scale Genomics, Transcriptomics, and Metagenomics for All

Philip Blood
Pittsburgh Supercomputing Center
blood@psc.edu

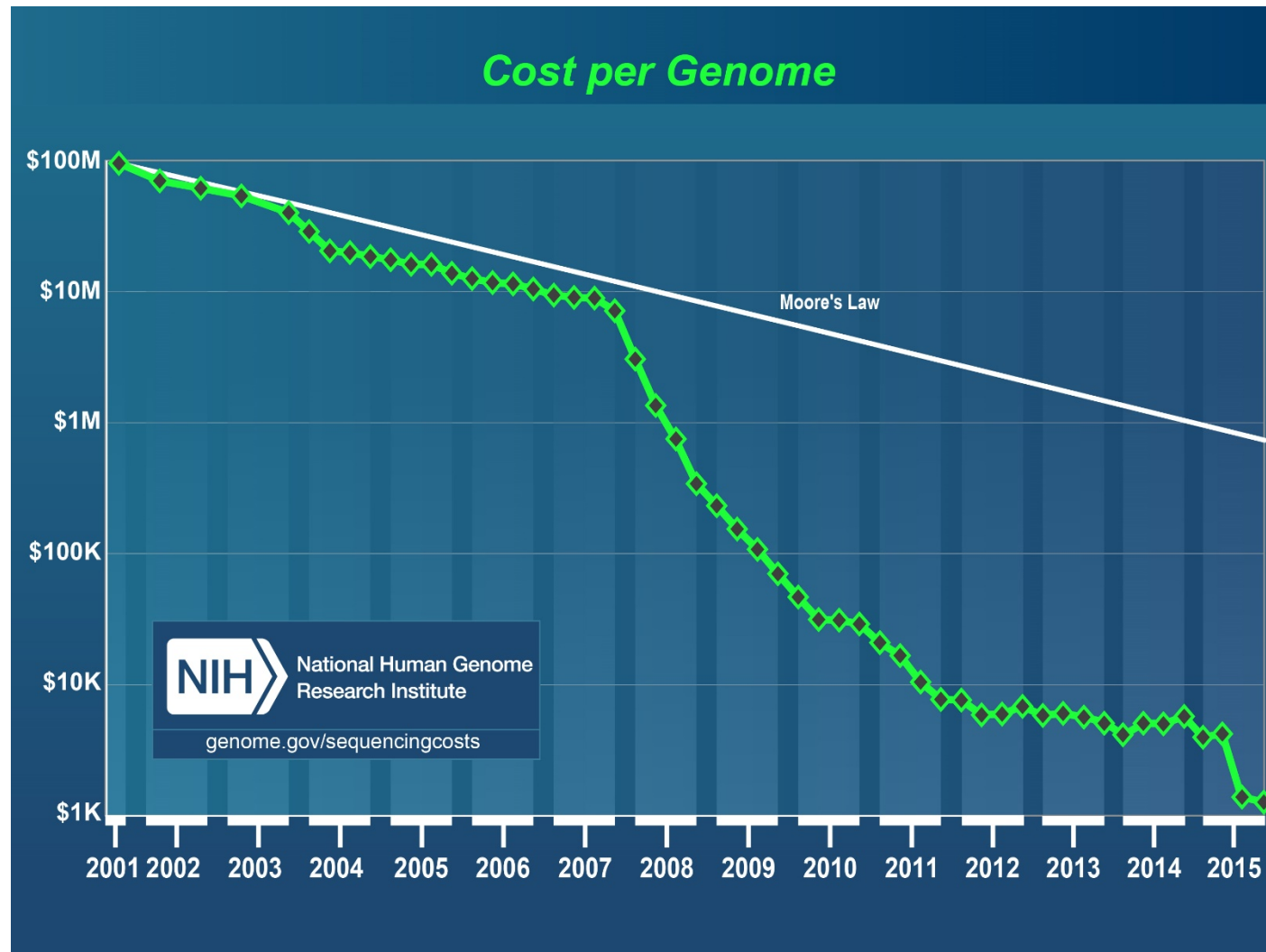
Genomics, Metagenomics, Transcriptomics

- All DNA in a cell is the genome
- DNA is transcribed into corresponding molecules of RNA
- The transcriptome is all of the RNA transcripts of a particular cell



Courtesy Lenore Pipes, Cornell

The obligatory genomics big data graph



https://www.genome.gov/images/content/costpergenome2015_4.jpg

Facilitating Genomics

- Substantial and growing need for bioinformatics help
- Various resources available: XSEDE, CyVerse, NCGAS, Galaxy, GenePattern, etc...
- Key Challenges:
 - Data wrangling
 - Choosing best software tools/ bioinformatics know-how
 - Utilizing HPC systems needed for large analyses
 - Reproducibility

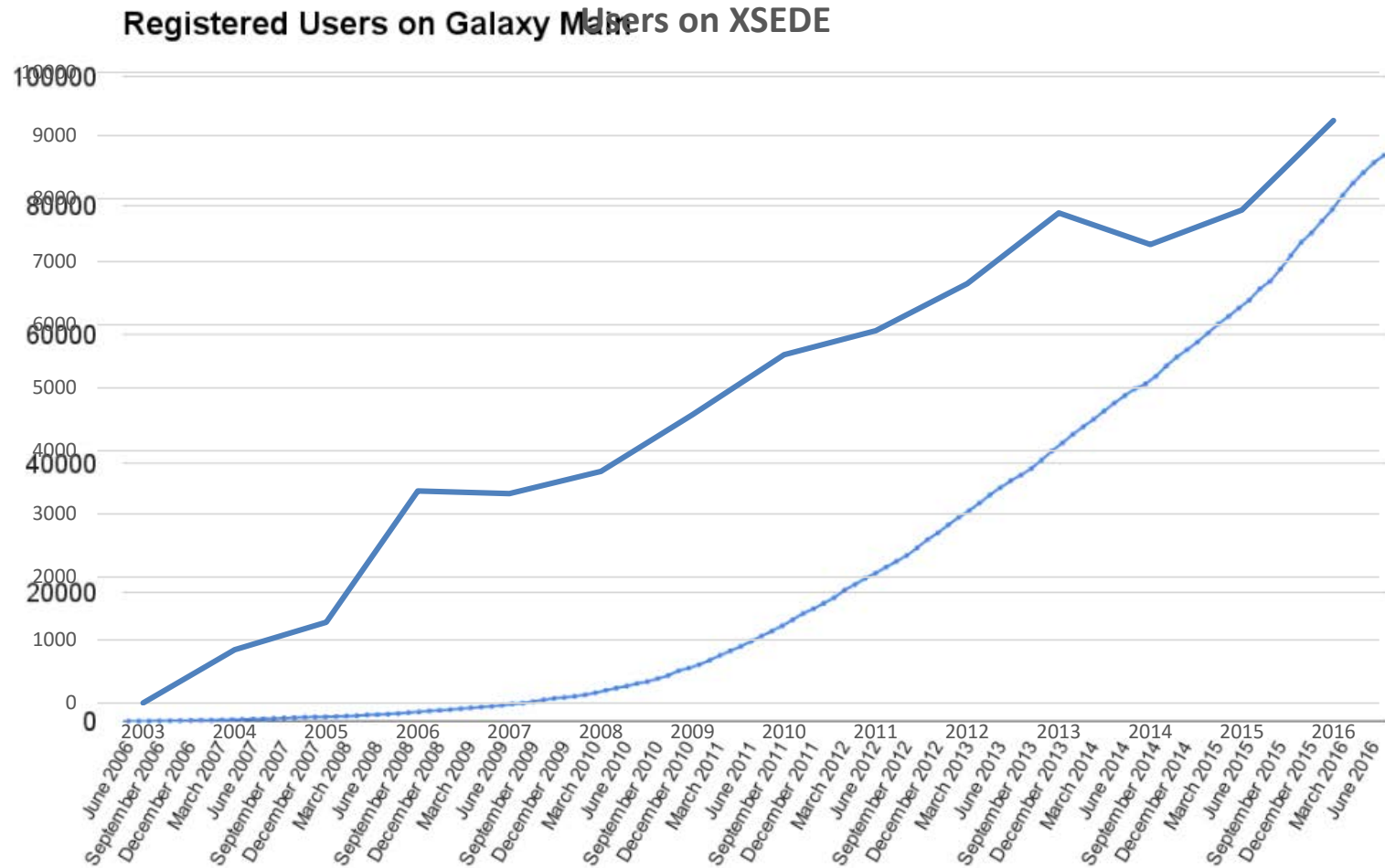
Installing Genomics Software!

hmmer/3.1b2	prodigal/2.6.3	abyss/1.9.0	htseq/0.6.1	icc/16.0.2)	python/2.7.11_gcc
annovar/2016.02.01	idba-tran/1.1.1	ansys/17.1	idba-tran/1.1.1_long	anvio/2.0.2	idba-ud/1.1.1
augustus/3.2.2	bamtools/2.4.0	bcftools/0.1.19	R/3.3.1-mkl	bcftools/1.3.1	
kallisto/0.43.0	raxml/8.2.9	bedops/2.4.19	ray/2.3.1	bedtools/2.25.0	repeatmasker/4.0.6
blasr/1.3.1	rnammer/1.2	blast/2.2.31	rsem/1.2.21	blat/v35	macs/1.4.3
sailfish/0.9.2	macse/1.2	salmon/0.6.0	mafft/7.300	salmon/0.7.2	samtools/0.1.19
samtools/1.3	malt/0.3.8	bowtie/1.1.2	masurca/3.1.3	bowtie2/2.2.7	matlab/MCR_R2013a
scythe/0.981	bwa/0.7.13	seqtk/1.2-r94	maxbin/2.1.1	sickle/1.33	megane/5.11.3
	canu/1.3	signalp/4.1c	cdbfasta/2013	mothur/1.38.1	cd-hit/2016.06.21
snvmix/0.11.8-r5					
soapdenovo2/2015-10-09	somaticsniper/1.0.5	spades/3.8.1	sra-toolkit/2.5.7	cufflinks/2.2.1	
strelka/1.0.14	dammit/0.3	mummer/3.23		deeptools/2.3.5	detonate/1.10
diamond/0.7.11	discover/52488)	discoverdenovo/52488		ectools/2014-12-01	
ngscheckmate/2016.10.12	tmhmm/2.0c	emboss/6.6.0	tophat/2.1.1	trimmomatic/0.36	falcon/0.4.1
openslide/3.4.1	trinity/2.0.6	fasta-splitter/0.2.4	paml/4.9a	trinity/2.1.1	fastqc/0.11.3
trinity/2.2.0	fastq-splitter/0.1.2	trinotate/2.0.2	fastx/0.0.14	trinotate_db/2.0	
trinotate_db/2.0_pylon1	pbbjelly/15.8.24	flash/1.2.11	perl/5.18.4-threads	unceqr/2016-07-08	
fraggenescan/1.20	varscan/2.4.2	vcftools/0.1.15	gatk/3.5	velvet/1.2.10-maxk63-big	
gatk/3.6	velvet/1.2.10-maxk63-categ14-big	phylosift/1.0.1		picard/2.1.1	
	pilon/1.16				
platanus/1.2.4	wgs/8.2	genome-music/0.4.1	plinkseq/0.10	wgs/8.3rc	xhmm/1.0
primer3/1.1.4		primer3/2.2.3	hisat2/2.0.4	primer3/2.3.7	hmmer/2.3.2
prodigal/2.6.2					

A Different Flavor of XSEDE ECSS

- Generally not interested in intensive optimization of a single code -- there are too many, and constantly changing!
 - Some major codes have been addressed through ECSS Community Codes (e.g. Trinity)
- Generally want to know:
 - What tools are best for the job?
 - Where can I run them?
 - How do I run them?
 - How do I write proposals for allocations on XSEDE?
- Engaging users through:
 - ECSS Novel and Innovative Projects
 - “Light” ESRT projects

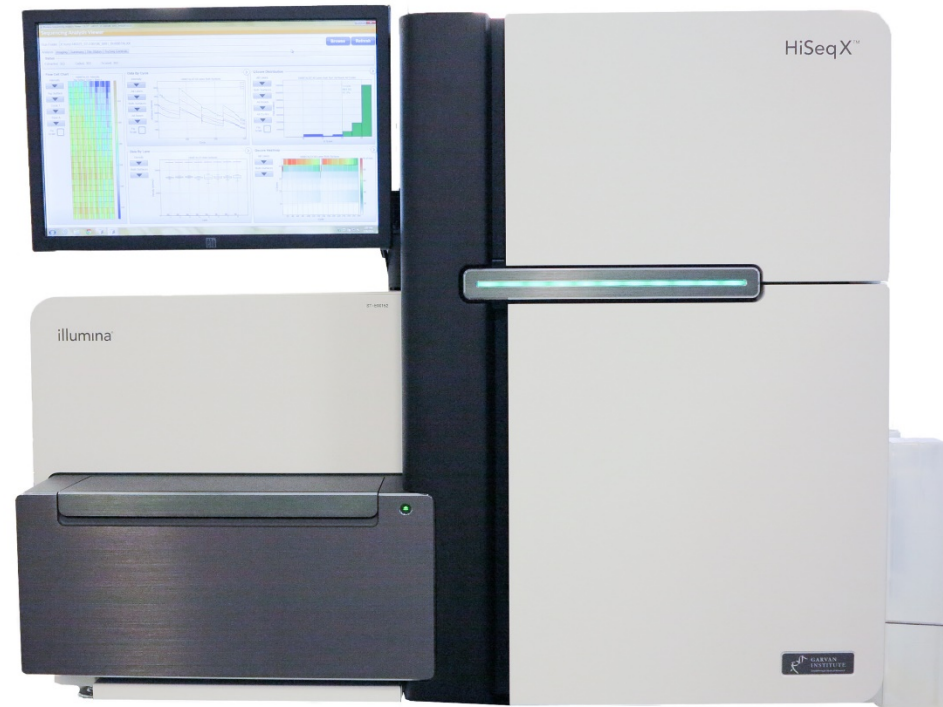
Genomics Community: Well-established in XSEDE?



Genome Assembly Using Next-generation Sequencing

- Reconstruct genomes of millions to billions of nucleotide base pairs (bp)
- ...containing repetitive sequences thousands of bp long
- ...using random 100-250 bp fragments (reads)*
- ...which have systematic and random errors
- Doing this reliably requires **deep coverage** and often **large shared memory**

*Bigger fragments can be generated, but at higher cost

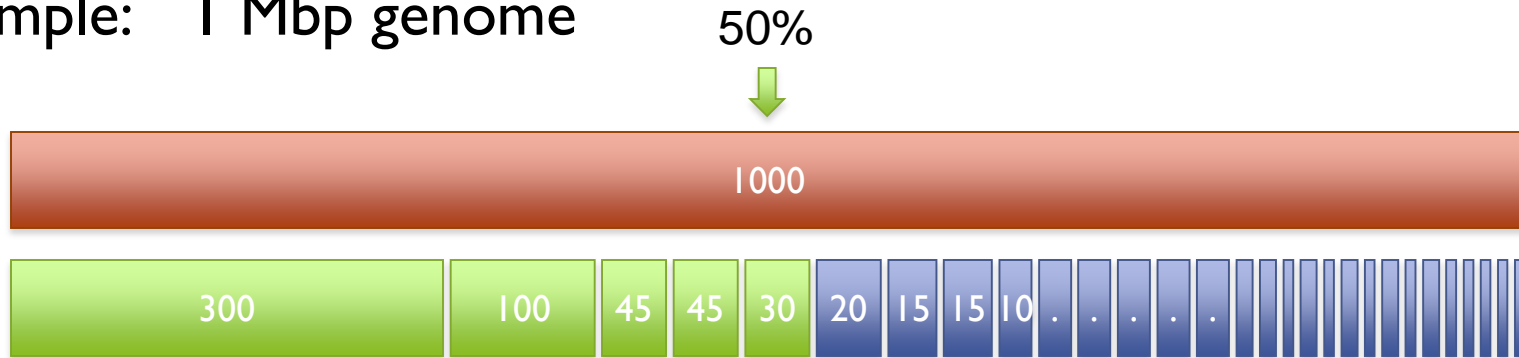


Kinghorn Centre for Clinical Genomics, Garvan
Institute of Medical Research.
Image credit: P. Morris/Garvan

N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp

$(300k + 100k + 45k + 45k + 30k = 520k \geq 500kbp)$

Note:

A “good” N50 size is a moving target relative to other recent publications. 10-20kbp contig N50 is currently a typical value for most “simple” genomes.

Slide courtesy of Michael Schatz, CSHL

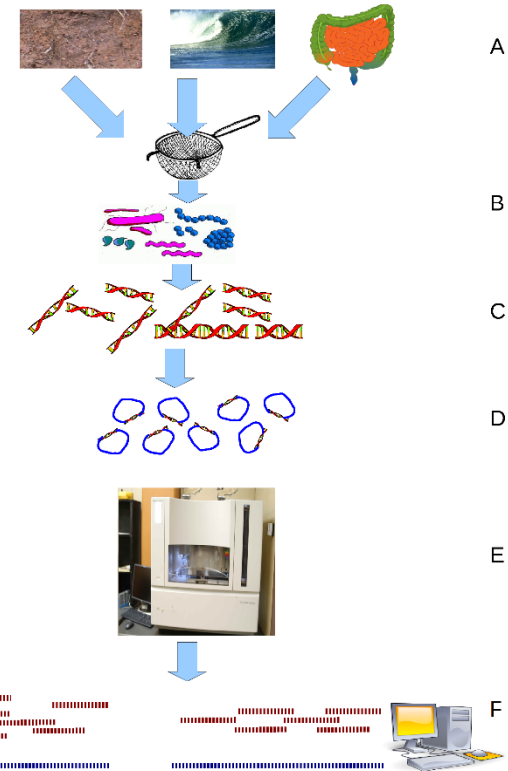
Improved SNP Detection in Metagenomic Populations

Ping Ma and Xin Xing, University of Georgia

- **Goal:** Develop statistical method that can distinguish closely related, unknown species in a metagenomic sample
- **Problem:** Didn't know where to start with analyzing large metagenomic data sets

★ ECSS Support

- Identified Ray MPI-based genome assembler capable of assembling large metagenomics data sets
- Tested Ray on *Bridges* to guide user on core counts for massive metagenome assemblies
- Helped user parallelize their R-based tool (MetaGen) on *Bridges*
- Helped user distribute data parallel jobs across many Bridges nodes using Slurm



Environmental Shotgun Sequencing (ESS). (A) Sampling from habitat; (B) filtering particles, typically by size; (C) DNA extraction and lysis; (D) cloning and library; (E) sequence the clones; (F) sequence assembly. By John C. Wooley, Adam Godzik, Iddo Friedberg - <http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000667>, CC BY 2.5, <https://commons.wikimedia.org/w/index.php?curid=17664682>

Improved SNP Detection in Metagenomic Populations

Ping Ma and Xin Xing, University of Georgia

- ✓ **Assembled 900 Gigabase pairs (Gbp) of gut microbial DNA from normal and diseased patients**
 - Created metagenomes for both type-2 diabetes (T2D) and inflammatory bowel disease (IBD) patients
 - Identified important pathogenic or missing probiotic species
 - Massive metagenome assemblies took only hours using an MPI-based metagenome assembler, Ray, on dozens of *Bridges* RM nodes connected by Omnipath
- ✓ **Improved characterization of composition of human gut microbiome**
 - First use of unsupervised binning method provides more accurate estimate of number of microbial species (~2000)
 - Estimated abundance of known vs. unknown species
 - MetaGen binning software ran across 10 RM nodes, clustering 500,000 contiguous sequences in only 24 hours
- ✓ **Identified important pathogenic or missing probiotic species in diseased patients**
 - Eight pathogenic microbial species identified in IBD patients
 - Two probiotic strains identified with lower abundance in T2D patients
 - Working to predict genes from unknown microbial species and identify genes related to disease conditions

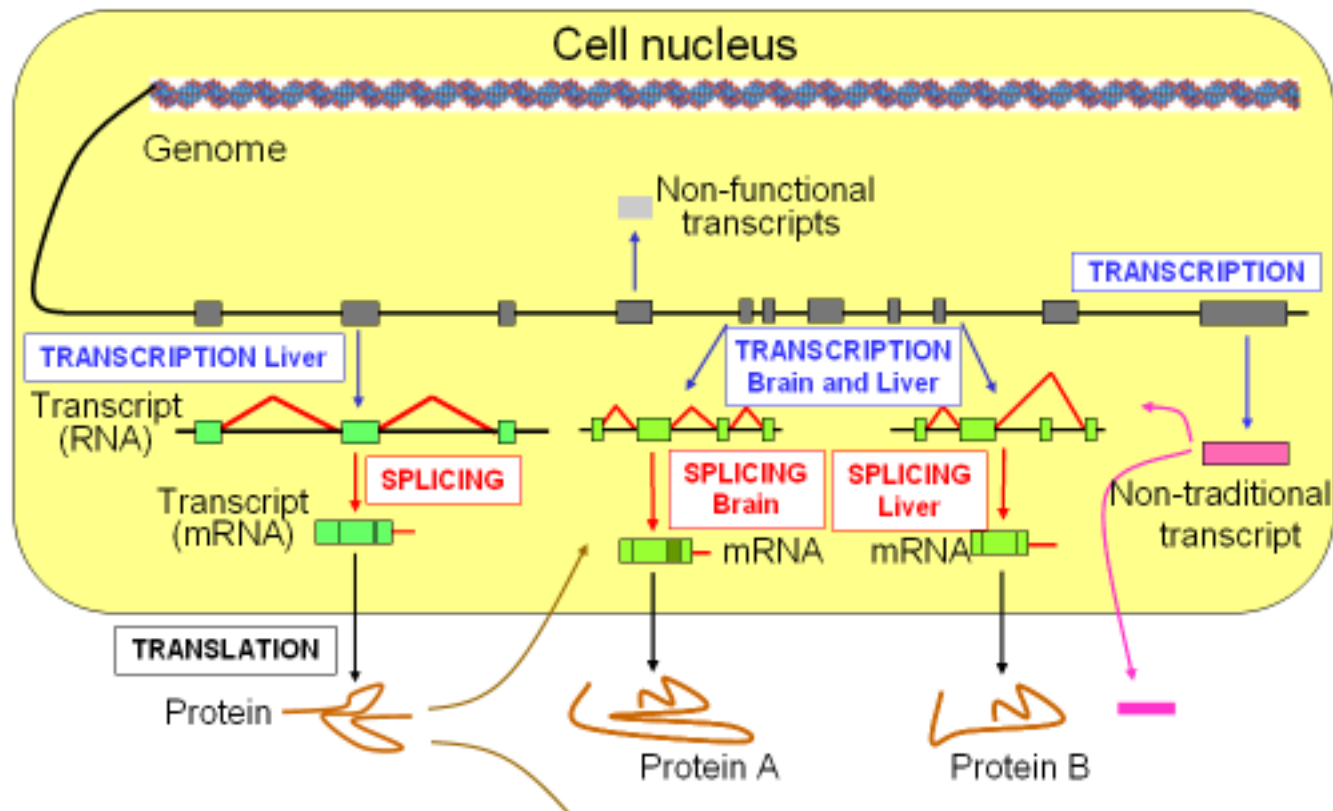
Factors affecting quality of transcriptome analysis

Raminder Singh
Indiana University

Jordi Abante, Noushin Ghaffari and Charles D. Johnson
Texas A&M

What can the transcriptome tell us?

- The transcriptome shows when and where each gene is turned off or on in the cells and tissues of an organism
- Counting the number of transcripts for a given gene can determine gene expression



Courtesy Lenore Pipes, Cornell

SEQC RNA-Seq Data

- Sequence Quality Control (SEQC) Consortium
- Six sites generated RNA-Seq data from well-studied human samples
- Coordinated by US Food and Drug Administration
- Sample A: Ten pooled cancer cell lines
- Sample B: Multiple brain regions from 23 donors

Goal: Determine Best Practices of Transcriptome Assembly

- Use a well-defined, well-understood standard data set
- Use a state-of-the-art transcriptome assembly and quality assessment pipeline
- Explore factors that influence quality of assembly
- This work is ongoing

Factors Influencing Assembly Quality

- Some factors to consider:
 - Sample preparation sites/methods (Aspects 1 & 3)
 - Sequencing depths (Aspect 2)
 - Preprocessing measures (Aspect 4)
 - Potentially many others
- So far have run 2 studies to completion:
 - Aspect 1: Compare outcomes across all six SEQC sites (site-effect)
 - Aspect 3: Compare sequencing library quality of SEQC sites vs. vendor-prepared library (library-effect)

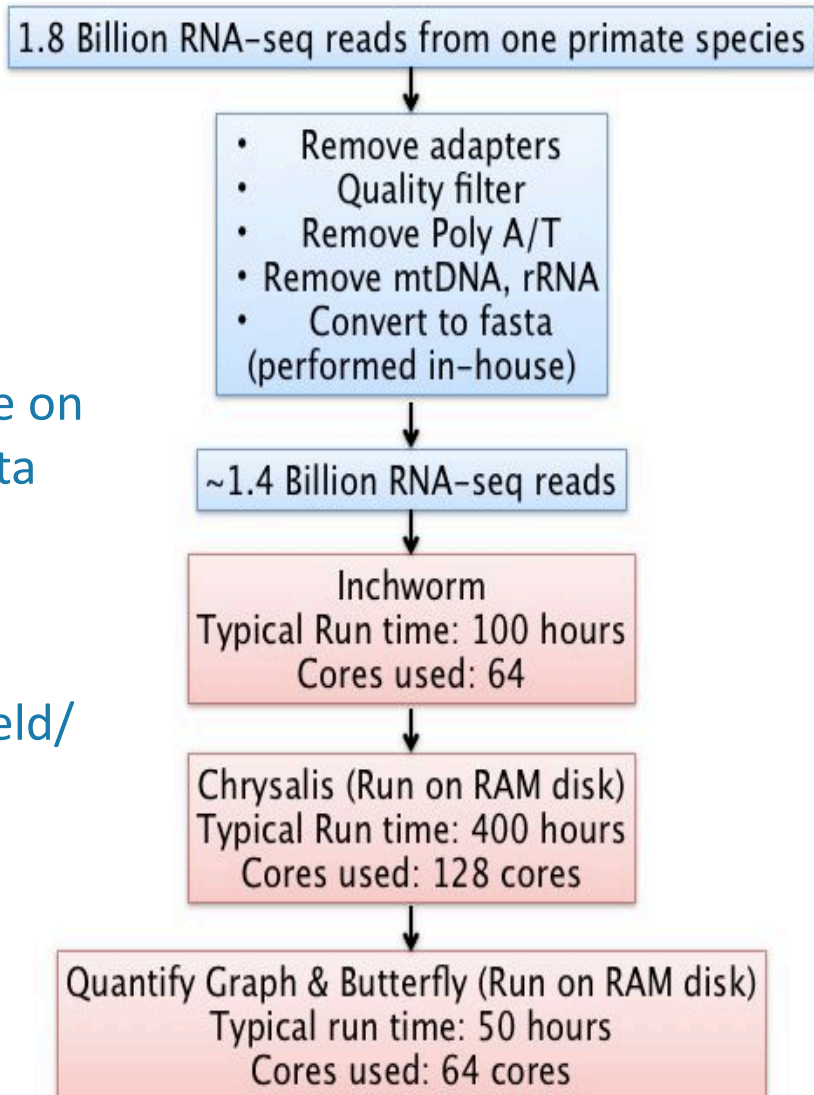
Workflow

- Pre-processing
 - Remove sequencing adapters, repetitive “tails”, rRNA and mitochondrial RNA, read error correction
- Assembly
 - Trinity transcriptome assembly pipeline
- Assembly Post Postprocessing and Quality Assessment
 - Statistical measures: N50
 - Assess completeness of transcriptome gene content
 - Map/Compare assembled transcriptome and original reads to reference genome
- **Problem:** Wrangling massive data sets through complex pipeline

Massive Transcriptome Assemblies with Trinity

Challenges:

- Run Trinity pipeline on unprecedented data set
- Optimized for Blacklight/Greenfield/Bridges



Trinity + files on
RAM disk required
~1 TB RAM

On Bridges, use
local disk

5x faster on
RAM/local
disk

Lots o' tools!

Tools used at different stages of pipeline

	Process/Tool	Purpose	CPU	Memory	Time(min.)	Dependencies
Pre-Processing	Cutadapt	Get rid of adapters.	16 CPUs	8 GB /CPU	50	python/2.7.9
	Flexbar	Get rid of poly A/T tails.	15 CPUs	8 GB /CPU	90	samtools/1.2, flexbar/2.5
	Bowtie	Remove reads coming from rRNA and chrM.	15 CPUs	8 GB /CPU	90	samtools/1.2, bowtie/1.1.1
	SEECER	Correct errors in reads to improve transcriptome assembly.	15 CPUs	24 GB/CPU	60	seecer/0.1.3
Reads Post-Processing	Tophat	Map the reads to the reference genome.	15 CPUs	8 GB/CPU	2040	tophat/2.1.0-all
	Genome coverage	Get coverage Tophat output.	4 CPUs	4 GB /CPU	20	Samtools
	Samtools sort	Sort bam kariotypically.	5 CPUs	4 GB/CPU	250	Samtools
	CreateSequenceDictionary	Generate dictionary from reference.	1 CPU	2 GB/CPU	10	Picard
	GATK	Call SNPs.	4 CPUs	4 GB /CPU	240	Java, Samtools, Picard
Assembly and Assembly Post-Processing	Trinity2	Assemble transcriptomes.	30 CPUs	66 GB/CPU	1500	trinity/2.0.6-all
	BUSCO	Assess transcriptome completeness with single-copy orthologs.	8 CPUs	2 GB/CPU	80	emboss, hmmer, ncbi-blast, python
	DETONATE	Evaluate transcriptome assemblies.	15 CPUs	8 GB/CPU	4320	detonate, bowtie2
	gmap_build	Builds a gmap database for the reference genome.	4 CPUs	4 GB /CPU	20	gmap
	gmap	Map the contigs to the reference genome	8 CPUs	4 GB/CPU	60	gmap
	Genome coverage	Samtools	4 CPUs	4 GB /CPU	20	Samtools
	Samtools sort	Sort bam kariotypically.	1 CPU	2 GB/CPU	300	Samtools
	Samtools index	Generate indexes (.bai)	1 CPU	2 GB/CPU	10	Samtools
	CreateSequenceDictionary	Generate dictionary from reference.	1 CPU	2 GB/CPU	10	Picard
	Samtools faidx	Index reference.	1 CPU	2 GB/CPU	2	Samtools
	GATK	call single-nucleotide polymorphisms (SNPs).	4 CPUs	4 GB /CPU	240	Java, Samtools, Picard

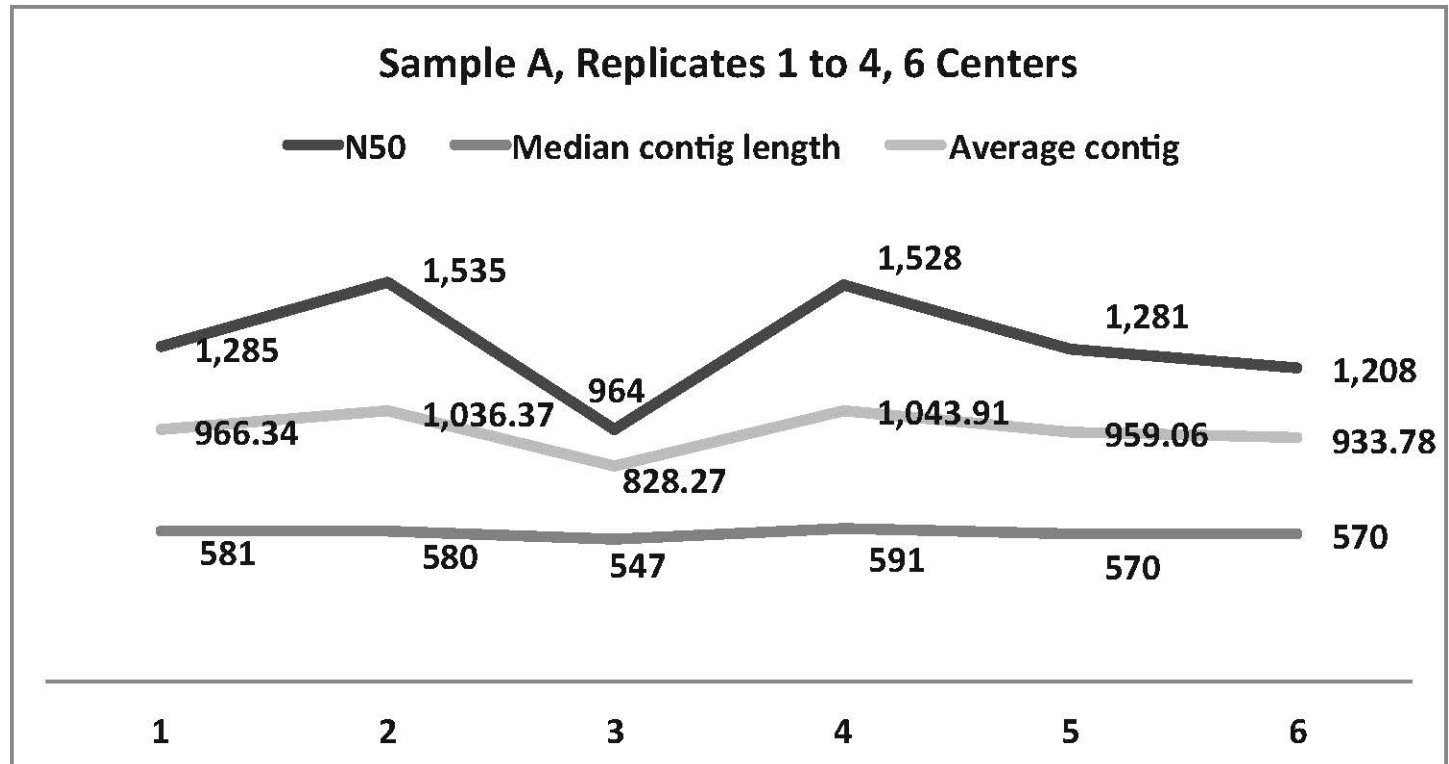
ECSS Support: Wrangling Data Through Pipelines

- *Computational Heterogeneity*
 - Lots of different tools with different requirements
 - large memory thread parallel
 - regular memory thread parallel
 - small single core jobs (process placement)
 - Significant effort required to run each tool optimally
 - **Workflow tools challenged by this heterogeneity**
- *Checkpoint and Recovery*
 - Long-running processes (days to weeks!)
 - Periodically take snapshots, especially when using local disk that is purged upon job failure/completion
 - To do: look at automated checkpointing

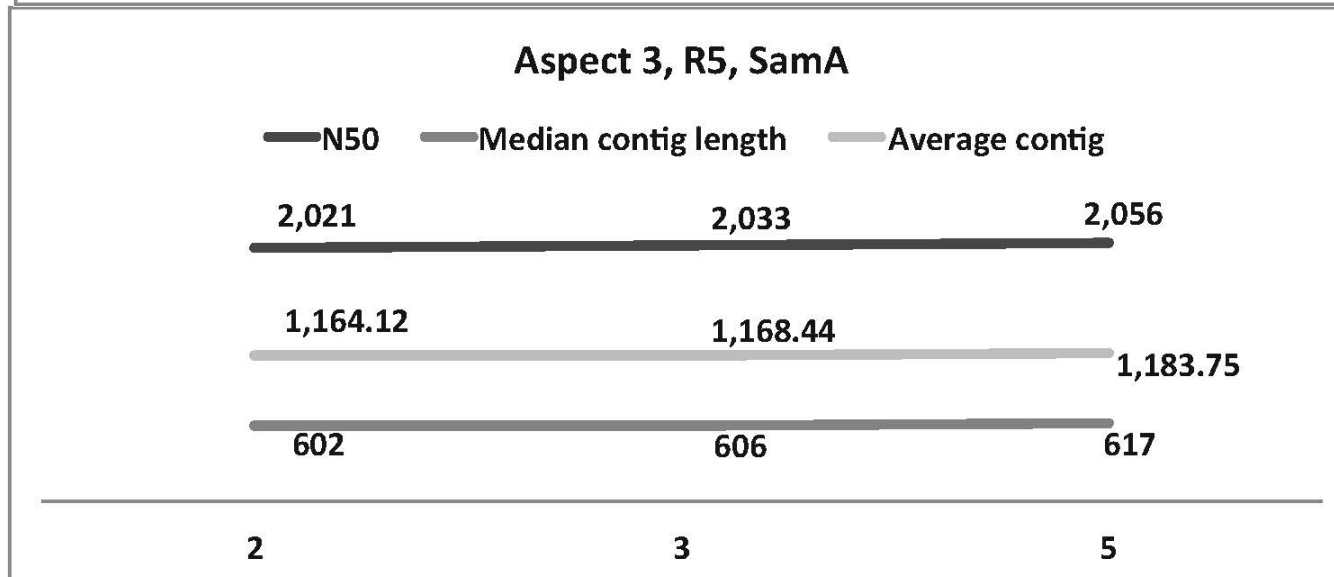
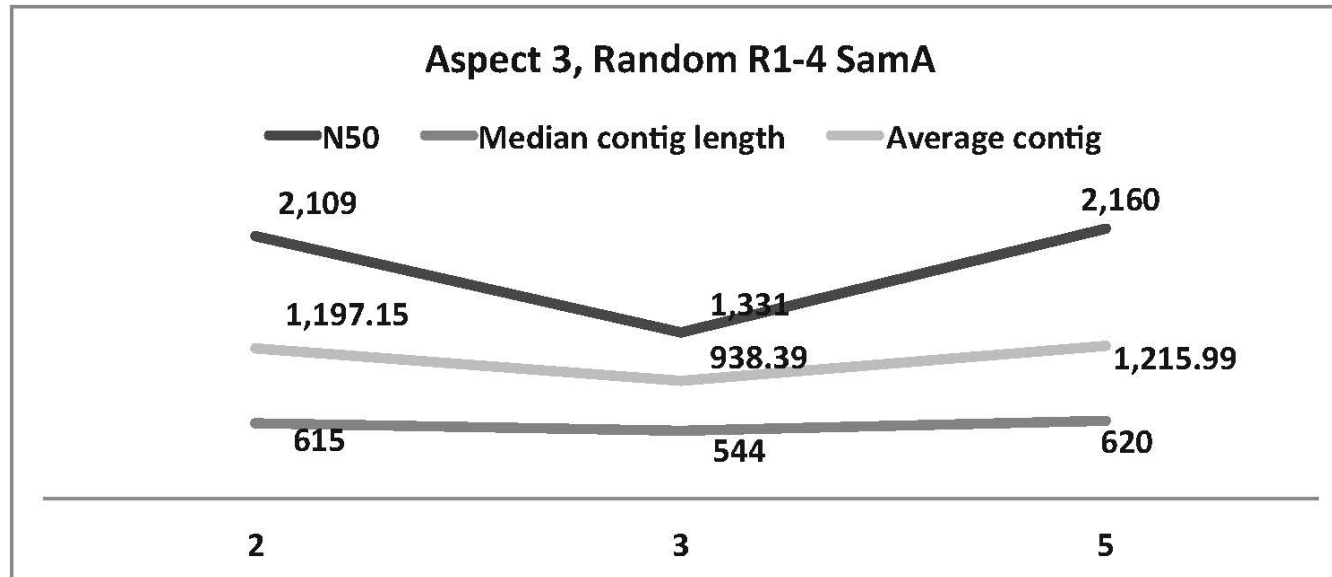
ECSS Support: Wrangling Data Through Pipelines

- *Data movement/management*
 - Local disk/ RAM disk needed for performance
 - Scratch → Local disk → Long term storage
- *Data verification*
 - Lots of data movement steps, with lots of files
 - Verify at each step or confusing issues arise downstream

Effect of Sequencing Site



Effect of Library Prep



Ongoing work

- Address other aspects affecting transcriptome quality
 - Sequencing depths (Aspect 2)
 - Preprocessing measures (Aspect 4)
 - Potentially many others (different tools, tool options, etc.)
- Make transcriptome analysis pipeline available through a gateway so others can benefit

Galaxy XSEDE Gateway

James Taylor, Johns Hopkins
Anton Nekrutenko and Nate Coraor, Penn State

- **Goal:** Enable large jobs on Galaxy Main to run on XSEDE
- **Problem:** Remote job submission in Galaxy; optimized tools for XSEDE

★ ECSS Support

- Helped with process of becoming a gateway and writing Gateway proposals
- Jobs being submitted to Stampede for over a year
- Created an optimized Trinity Galaxy tool to run on Bridges - now active!

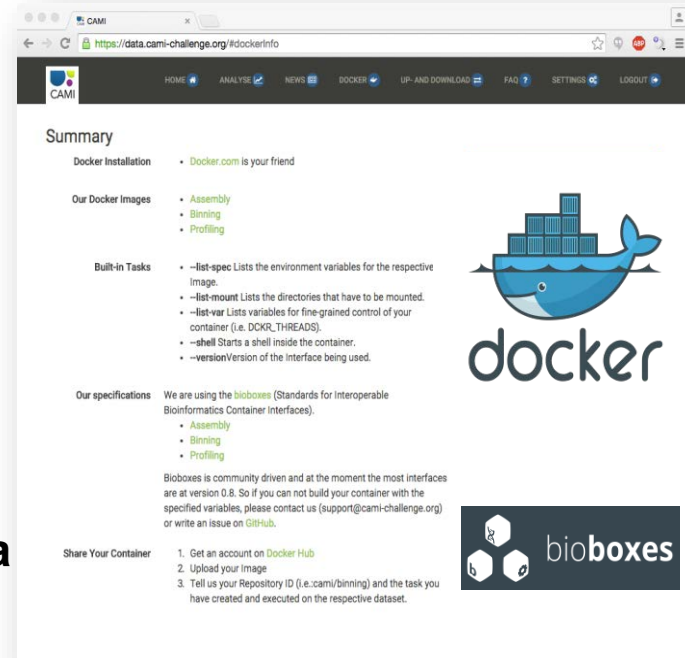
The screenshot displays the Galaxy XSEDE Gateway interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', 'User', and a 'Using 0%' status indicator. The left sidebar lists various tools and categories, including 'NGS: RNA Structure', 'NGS: Du Novo', 'NGS: Gemini', 'NGS: Assembly', 'Trinity (Beta) De novo assembly of RNA-Seq data Using Trinity on PSC's Bridges', 'Operate on Genomic Intervals', 'Statistics', 'Graph/Display Data', 'Phenotype Association', 'BEDTools', 'Genome Diversity', 'EMBOSS', 'Regional Variation', 'FASTA manipulation', 'Multiple Alignments', 'Metagenomic Analysis', 'Multiple regression', 'Multivariate Analysis', 'Motif Tools', 'STR-FM: Microsatellite Analysis', 'NCBI SRA Tools', 'DEPRECATED', 'NGS: GATK Tools (beta)', and 'CloudMap'. The main panel shows the configuration for the 'Trinity (Beta) De novo assembly of RNA-Seq data Using Trinity on PSC's Bridges' tool (Galaxy Version 0.0.1). The configuration includes a dropdown for 'Paired or Single-end data?' set to 'Paired', input fields for 'Left/Forward strand reads' and 'Right/Reverse strand reads' (both showing 'No fasta or fastq dataset available.'), a dropdown for 'Strand-specific Library Type' set to 'Not set', input fields for 'Group pairs distance' (500) and 'Path reinforcement distance' (75), a dropdown for 'Use Additional Params?' set to 'No', and a dropdown for 'Job Resource Parameters' set to 'Use default job resource parameters'. An 'Execute' button is at the bottom. The right sidebar shows the 'History' panel with a search bar and a list of datasets, including '1: UCSC Main on Human: knownGene (genome)'.



Critical Assessment of Metagenomic Interpretation

<http://www.cami-challenge.org>

- interpretation of metagenomes relies on computational approaches
 - **short read assembly**
 - **taxonomic binning/classification**
 - **taxonomic profiling**
- CAMI aims at independent, comprehensive and bias-free **evaluation of methods**
- extensive high-quality **unpublished metagenomic data sets**
- results will provide reproducible and quantitative **measurements of tool performance**
- will serve as
 - **guide** to users
 - help developers identify **directions for future work**

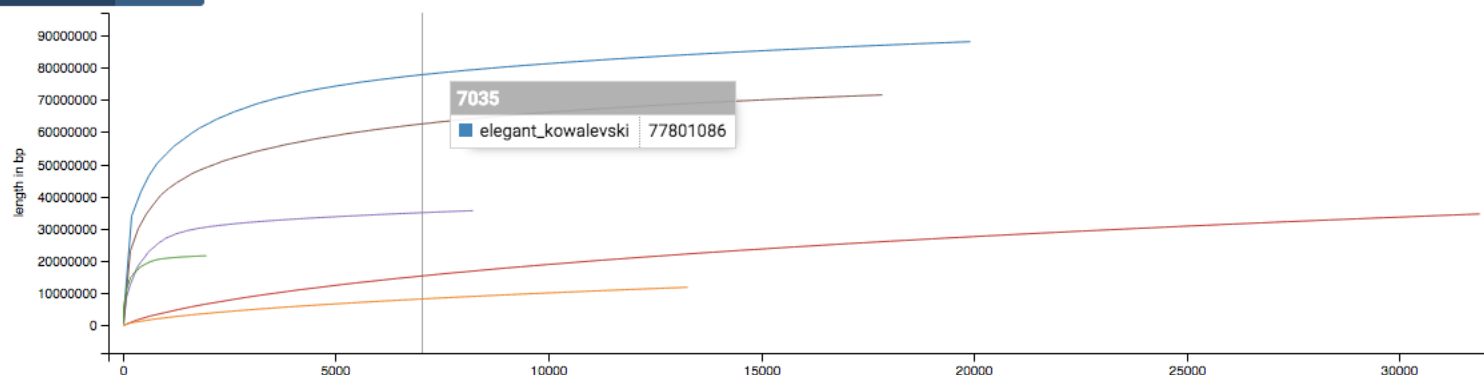


Follow @CAMI_challenge

CAMI Evaluation Metrics

Anonymous Name	# contigs	Largest contig	Total length	N50	GC (%)	# misassemblies	# relocations	# translocations	# inversions	# misassembled contigs	# indels
pensive_babbage	10.0	6503724.0	8361599.0	6503724.0	46.96	-	-	-	-	-	-
focused_bardeen	8864.0	487875.0	3.680766E7	23804.0	54.17	28.0	5.0	23.0	0.0	27.0	-
sharp_perلمان	17911.0	888870.0	7.155452E7	19216.0	54.92	-	-	-	-	-	-
adoring_jones	16018.0	457213.0	4.7181912E7	13601.0	54.0	-	-	-	-	-	-
goofy_darwin	15795.0	888811.0	4.8721356E7	28403.0	53.75	172.0	33.0	130.0	9.0	125.0	93.0
elegant_kowalevski	20004.0	2780101.0	8.807724E7	24752.0	54.62	0.0	0.0	0.0	0.0	0.0	-
trusting_colden	184.0	8.0	133.0	816.0	0.18	67.0	14.0	4775.0	9.77	-	-
lonely_franklin	32148.0	10239.0	3.4692516E7	1183.0	56.54	45.0	6.0	37.0	2.0	45.0	-
drunk_galileo	8218.0	307410.0	3.558718E7	23934.0	54.41	-	-	-	-	-	-
hungry_jones	13325.0	15618.0	1.1841911E7	889.0	55.27	0.0	0.0	0.0	0.0	0.0	4.0
elated_wright	1957.0	515047.0	2.1586394E7	65409.0	57.52	80.0	24.0	49.0	7.0	65.0	100.0

[CONTIG LENGTH](#)
[NX](#)
[GC](#)



20 Storage Building Blocks, implementing the parallel *Pylon* filesystem (~10PB) using PSC's SLASH2 filesystem

4 MDS nodes

2 front-end nodes

2 boot nodes

8 management nodes

6 "core" Intel OPA edge switches: fully interconnected, 2 links per switch

Intel OPA cables

4 ESM (12TB) compute nodes

2 gateways per ESM

42 LSM (3TB) compute nodes

12 database nodes

6 web server nodes

20 "leaf" Intel OPA edge switches

32 RSM nodes with NVIDIA next-generation GPUs

16 RSM nodes with NVIDIA K80 GPUs

800 RSM (128GB) compute nodes, 48 with GPUs

<https://www.psc.edu/index.php/bridges-virtual-tour>