University of North Carolina at Chapel Hill

Center for Bioinformatics

- Home
- About
- Personnel
- Resources
- Software
- Education
- Support
- Links

The staff at the Center for Bioinformatics promote the use of computational tools for molecular biology, genetics, protein chemistry, and biochemistry research at the University of North Carolina.

The center offers in-depth workshops on topics ranging from DNA, RNA and protein sequence analysis to database searching, genomic predictions and molecular modeling.

http://bioinformatics.unc.edu

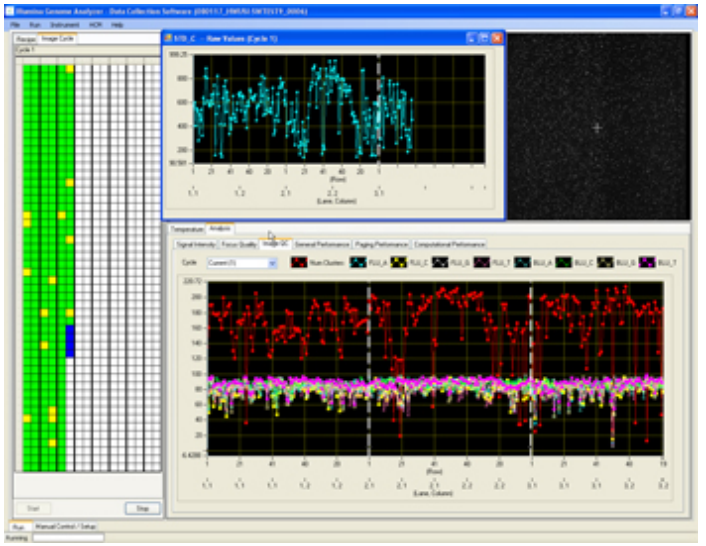
HT-SEQ at UNC

- Significant investment
- Started with an Illumina Genome Analyzer I (GAI) in November 2007 – will be upgraded to a GAII in place
- Added a second Illumina Genome Analyzer II (GAII) in April 2008
- Started doing “paired-end” runs
- Will shortly add a Roche 454

Illumina includes

- GA I - attached data collection workstation (windows)
- No firewalls and/or other software (default recommendation)
- Can store only one run in default configuration supplied (1 TB internal disk)
- Data processing software - Illumina Pipeline (currently v.0.3)
- If starting now – likely get the IPAR upgrade

Integrated Primary Analysis and Reporting (IPAR)



- Hardware and Software solution
- real-time image analysis and intensity calculations
- Software is *not* ready as of now

Integrated Primary Analysis and Reporting (IPAR)

- 4 x Quad core (16 cores total) HP DL560 server with a 2TB direct attached storage array
- UPS included
- Runs 64-bit windows
- Performs as advertised?

Moving the data …

- Sequencer is controlled via USB
- Illumina supplies “robocopy” to facilitate data transfer
- Does *not* work with external USB drives
- Will likely work with NFS-exported shares

- “Sneaker-net”
- External USB hard-drives (1 TB) – conveniently fit one Illumina run
- Firewire enclosure – best bet
- Data computer has Firewire port available

IT Infrastructure - Minimal

- Data - source to processing
- If you can move the data directly to processing server .. Lucky you!
- Dell PowerEdge 2950 server – Dual quad-core Xeon (2.66 GHz, 8-cores), 32 GB RAM, 6 x 1 TB SATA drives in RAID 5, RedHat EL 4.0.
- MD1000 SAS direct attached disk array – 15 x 1 TB SATA disks
- Most important resource – People

Data Processing

- Copying a TB of data over USB 2.0 takes 10-12 hours
- Illumina flow cell – 8 lanes
- 8 cores ←→ 8 lanes
- Image analysis – uses Genome Analyzer Pipeline software
- Linux based suite of programs – Perl, Python scripts

Pipeline “modules”

Module	Name of the program	Function	Address
Goat	Firecrest	Image analysis	Pipeline/Goat/goat_pipeline.py
	Bustard	Base-caller	Pipeline/Goat/bustard.py
Gerald	Eland	Sequence alignment	Pipeline/Gerald/GERALD.pl

Analysis “Pipeline”

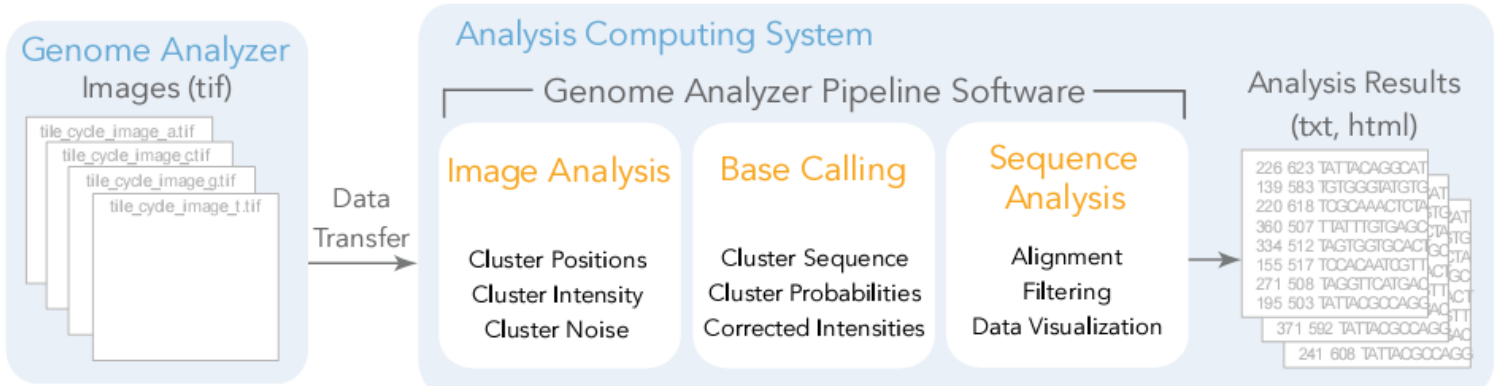


Figure 1 *Three Steps of Data Analysis*

- Data processing – 10-12 h for image analysis, 3 hours for base calling and 8 – 10 h for alignment followed by quality based filtering (human samples)
- Not as bad as it sounds
- Data processing is generally complete within 48 h after a run is complete

Deliverables

- "Efficient Local Alignment of Nucleotide Data" (ELAND) - written by Anthony J. Cox for Solexa
- Max 32 bp match – with 0,1,2 mismatches

- Raw sequence file – s_*_eland_query.txt
- Alignment results file – s_*_eland_multi.txt
- Sequence file – s_*_sequence.txt
- Export file – s_*_export.txt

Analysis and Visualization

- “in-house” bioinformatics resources in client lab
- “end user” access to the data

- Eland, Blat, MAQ (sanger), Soap, Mosaik
- EagleView, MaqView

- DNAStar genome assembler and Lasergene
- CLC Genomics workbench