Answering the demands of digital genomics

Michael Schatz

Sept 19, 2011

Beyond the Genome

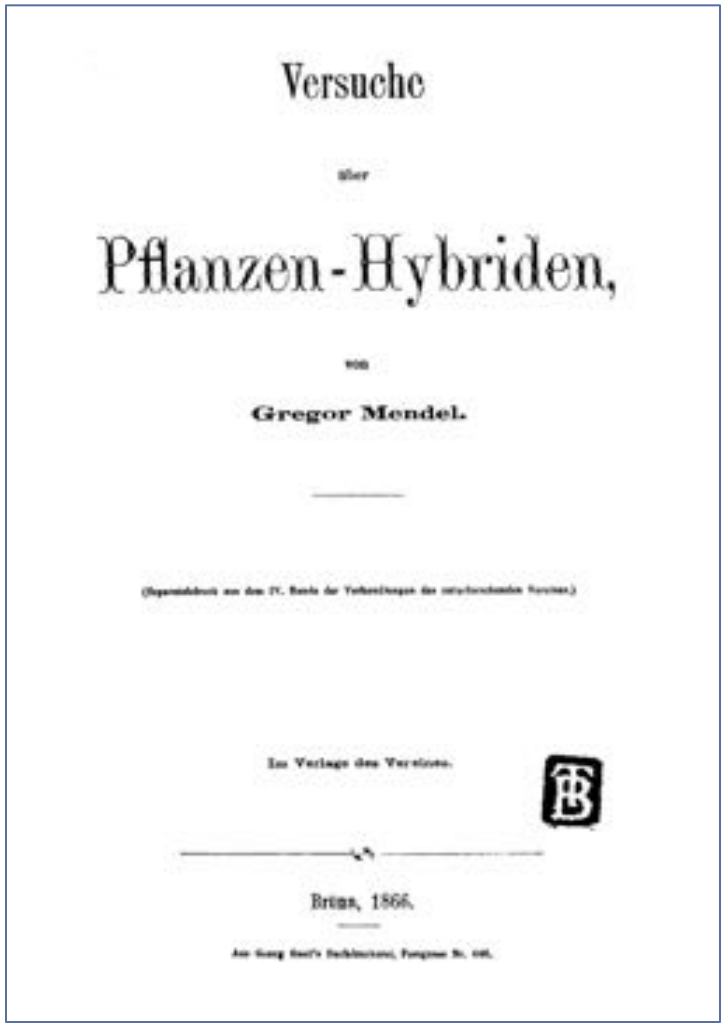
Outline

1. Milestones in genomics

2. The demands of genomics

3. Genomics in 2011 and beyond
    1. Hadoop and MapReduce
    2. Hadoop Applications for Genomics
    3. *Jnomics* case-study of esophageal cancer

Milestones in Genomics



Versuche

über

Pflanzen-Hybriden,

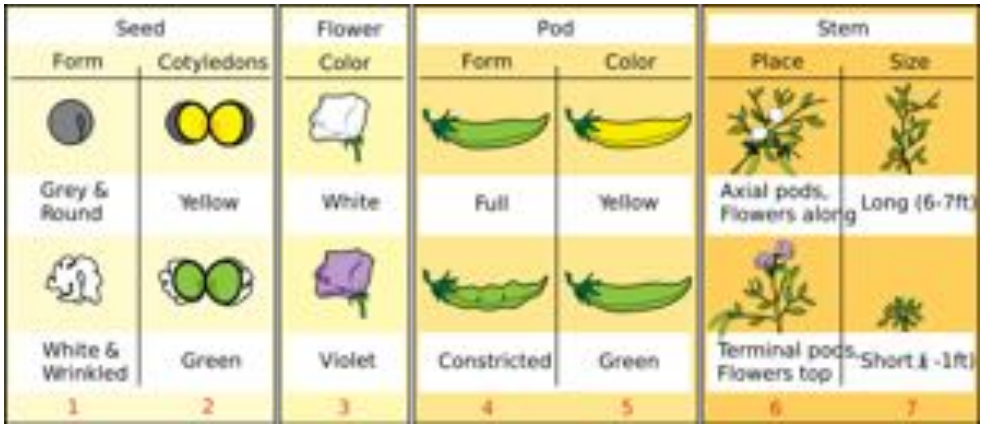
von

Gregor Mendel.

Brünn, 1866.

Observations of 29,000 pea plants and 7 traits

Generation	A	Aa	a	in Verhältniss gestellt: A : Aa : a
1	1	2	1	1 : 2 : 1
2	6	4	6	3 : 2 : 3
3	28	8	28	7 : 2 : 7
4	120	16	120	15 : 2 : 15
5	496	32	496	31 : 2 : 31
n				2^n-1 : 2 : 2^n-1

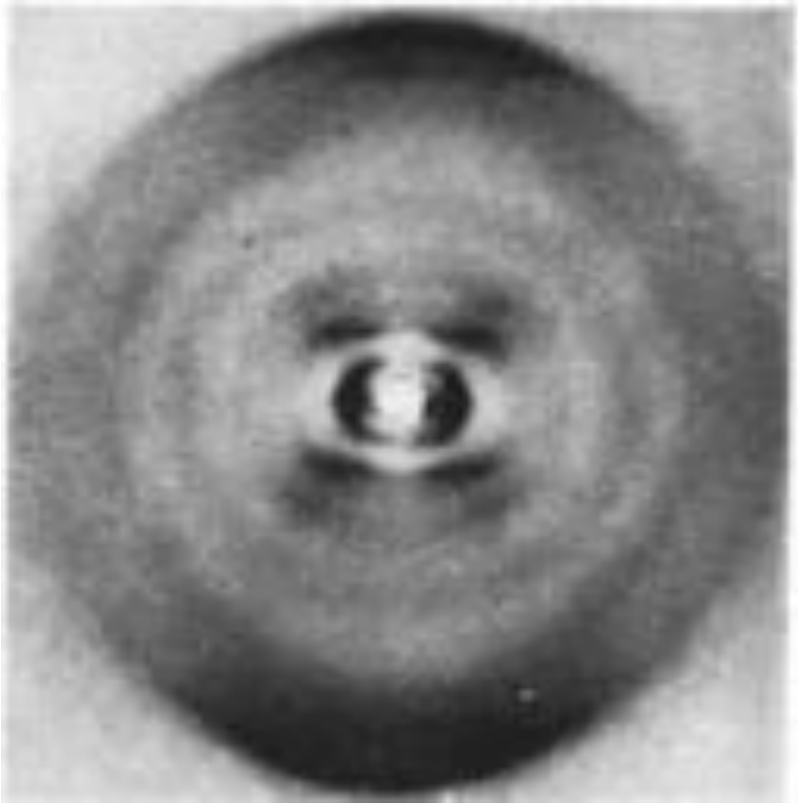


http://en.wikipedia.org/wiki/Experiments_on_Plant_Hybridization

Versuche über Pflanzen-Hybriden. Verh. Naturforsch (Experiments in Plant Hybridization)
Mendel, G. (1866). Ver. Brünn 4: 3–47 (in English in 1901, J. R. Hortic. Soc. 26: 1–32).

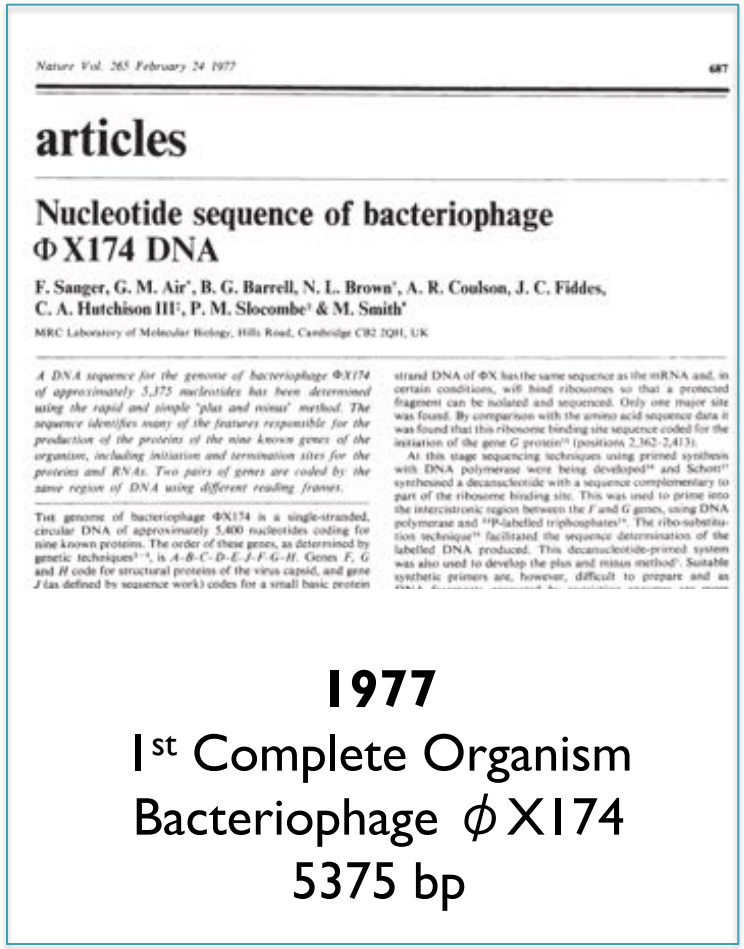
Milestones in Genomics

The origin and behavior of mutable loci in maize
McClintock, B (1950) *Proceedings of the National Academy of Sciences.* 36:344–55.



Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid
Watson JD, Crick FH (1953). *Nature* 171: 737–738.

Milestones in Genomics



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown*, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III*, P. M. Slocombe* & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[1-4], is *A-B-C-D-E-J-F-G-H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[5] (positions 2,362–2,413).

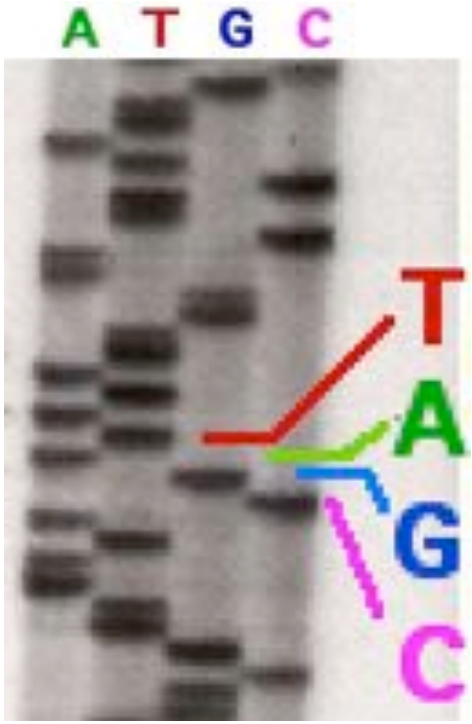
At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[6,7] and Schott[7] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *J* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[7]. The ribo-substitution technique[8] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[9]. Suitable synthetic primers are, however, difficult to prepare and as

1977

1st Complete Organism

Bacteriophage ϕX174

5375 bp



Radioactive Chain Termination

5000bp / week / person

http://en.wikipedia.org/wiki/File:Sequencing.jpg

http://www.answers.com/topic/automated-sequencer

Nucleotide sequence of bacteriophage ϕX174 DNA

Sanger, F. et al. (1977) *Nature. 265: 687 - 695*

Milestones in Genomics

1995
Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

2000
Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

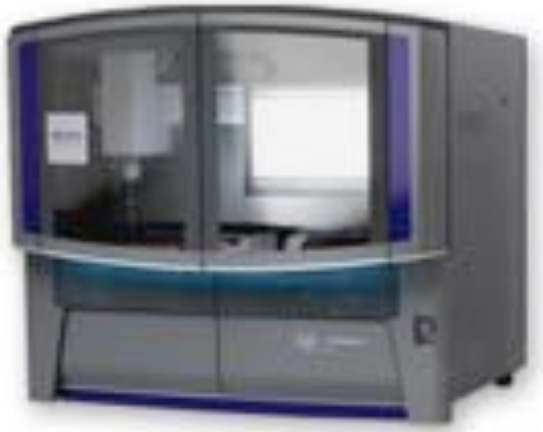
2001
Venter *et al.* / IHGSC
Human Genome
Celera Assembler. 2.9 Gbp

ABI 3700: 500 bp reads x 768 samples / day = 384,000 bp / day.
"The machine was so revolutionary that it could decode in a single day the same amount of genetic material that most DNA labs could produce in a year." J. Craig Venter

Milestones in Genomics

2004
454/Roche
Pyrosequencing
Current Specs (Titanium):
1M 400bp reads / run =
1Gbp / day

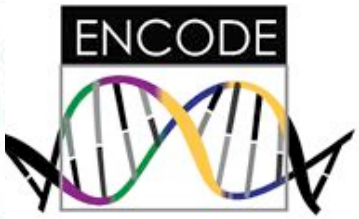
2007
Illumina
Sequencing by Synthesis
Current Specs (HiSeq 2000):
2.5B 100bp reads / run =
60Gbp / day



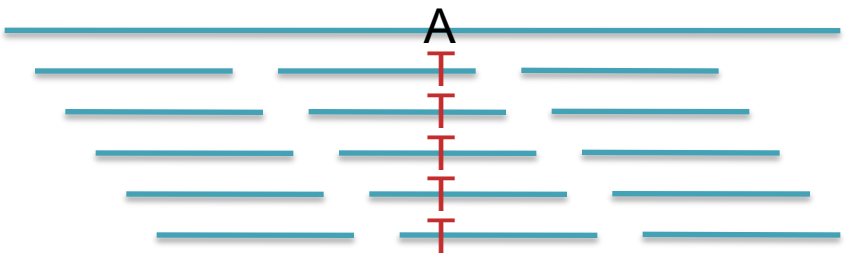
2008
ABI / Life Technologies
SOLiD Sequencing
Current Specs (5500xl):
5B 75bp reads / run =
30Gbp / day

Milestones in Genomics

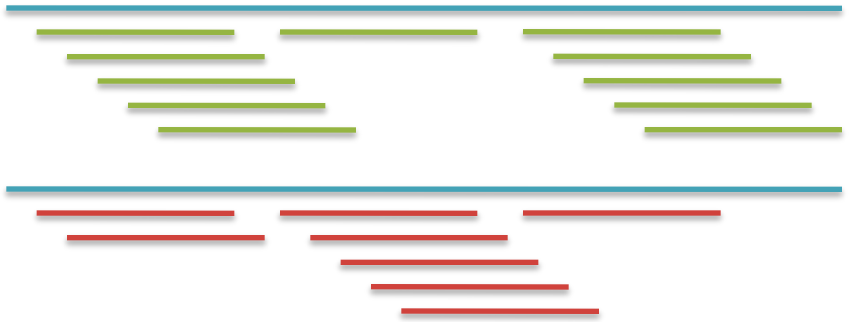




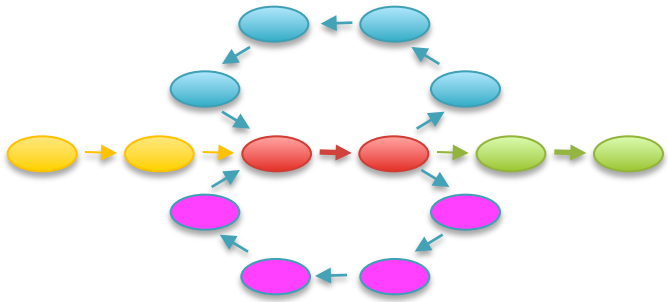
Alignment & Variations



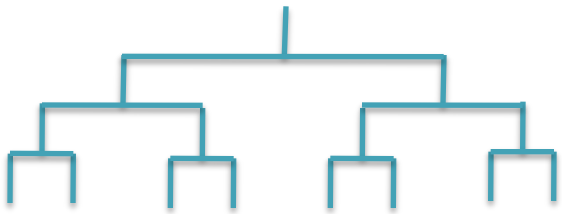
Differential Analysis



De novo Assembly



Phylogeny & Evolution

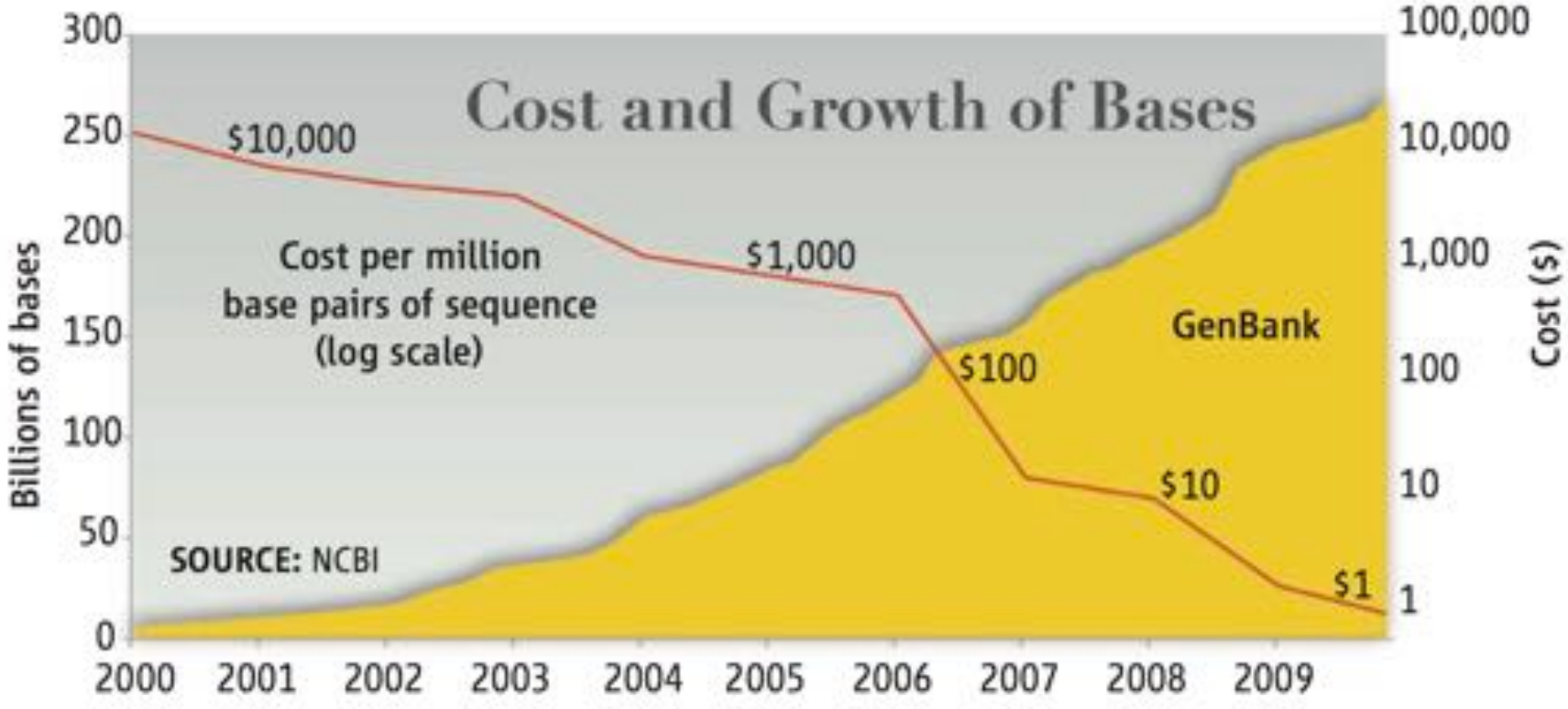


Sequencing Centers

Next Generation Genomics: World Map of High-throughput Sequencers
http://pathogenomics.bham.ac.uk/hts/

DNA Data Tsunami

Current world-wide sequencing capacity exceeds 13Pbp/year and is growing at 5x per year!



"Will Computers Crash Genomics?"
Elizabeth Pennisi (2011) *Science*. 331(6018): 666-668.

Beyond the Genome

- The cornerstones of genomics continue to be *observation*, *experimentation*, and *interpretation* of the living world
  - Technology has and will continue to push the frontiers of genomics
  - Measurements will be made *digitally* in great quantities, at extremely high resolution, and for diverse applications

- Demands of digital genomics
  1. *Experimental design*: selection, collection, tracking & metadata
     - Ontologies, LIMS, sample databases
  2. *Observation*: measurement, storage, transfer, computation
     - Algorithms to overcome sensor errors & limitations, computing at scale
  3. *Integration*: multiple samples, multiple assays, multiple analyses
     - Reproducible workflows, common formats, resource federation
  4. *Discovery*: visualizing, interpreting, modeling
     - Clustering, data reduction, trend analysis

Observational demands

- Overcome sensor/sequencing limitations through smarter algorithms
  - Co-development of protocol and computational methods
  - Can't sequence entire genomes -> Whole genome shotgun assembly
  - Reads have sequencing errors -> model error types, correct for them
  - Mate-pair protocols fail -> filter redundant pairs, failed mates
- Overcome computing limitations through parallel computing
  - Sensors improving faster than processors, using multiple processors at once
  - GNU Parallel is my new favorite command, limited by cores
  - Batch systems well established for embarrassingly parallel computation, limited by algs.
  - Hadoop, MPI, etc for more flexibility, limited by tools
- Overcome storage & transfer limitations through improved technology
  - Compress, filter, throw away
  - Transfer: Buy higher capacity internet, use smarter protocols
  - Storage: Buy higher capacity disk, parallel file systems, tiered storage

Hadoop MapReduce

http://hadoop.apache.org

- MapReduce is Google's framework for large data computations
  - Data and computations are spread over thousands of computers
    - Indexing the Internet, PageRank, Machine Learning, etc… (Dean and Ghemawat, 2004)
    - 946 PB processed in May 2010 (Jeff Dean at Stanford, 11.10.2010)
  - Hadoop is the leading open source implementation
    - Developed and used by Yahoo, Facebook, Twitter, Amazon, etc
    - GATK is an alternative implementation specifically for NGS

- Benefits
  - Scalable, Efficient, Reliable
  - Easy to Program
  - Runs on commodity computers

- Challenges
  - Redesigning / Retooling applications
    - Not Condor, Not MPI
    - Everything in MapReduce

Hadoop for NGS Analysis

Crossbow

Searching for SNPs with cloud computing

Identify 3.2M SNPs from 30x coverage in 4 hours for $85.

http://bowtie-bio.sf.net/crossbow/

(Langmead, Schatz, Lin, Pop, Salzberg, 2009)

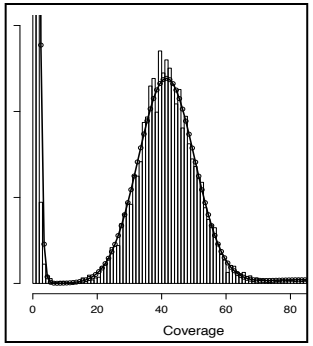
Contrail

Assembly of large genomes using cloud computing

Assemble the human genome on commodity computers with 24GB RAM

(Schatz, 2010)

http://contrail-bio.sf.net



Quake

Quality-aware error correction of short reads

Correct 97.9% of errors with 99.9% accuracy

http://www.cbcb.umd.edu/software/quake/

(Kelley, Schatz, Salzberg, 2010)

Genome Indexing

Rapid Parallel Construction of Genome Index

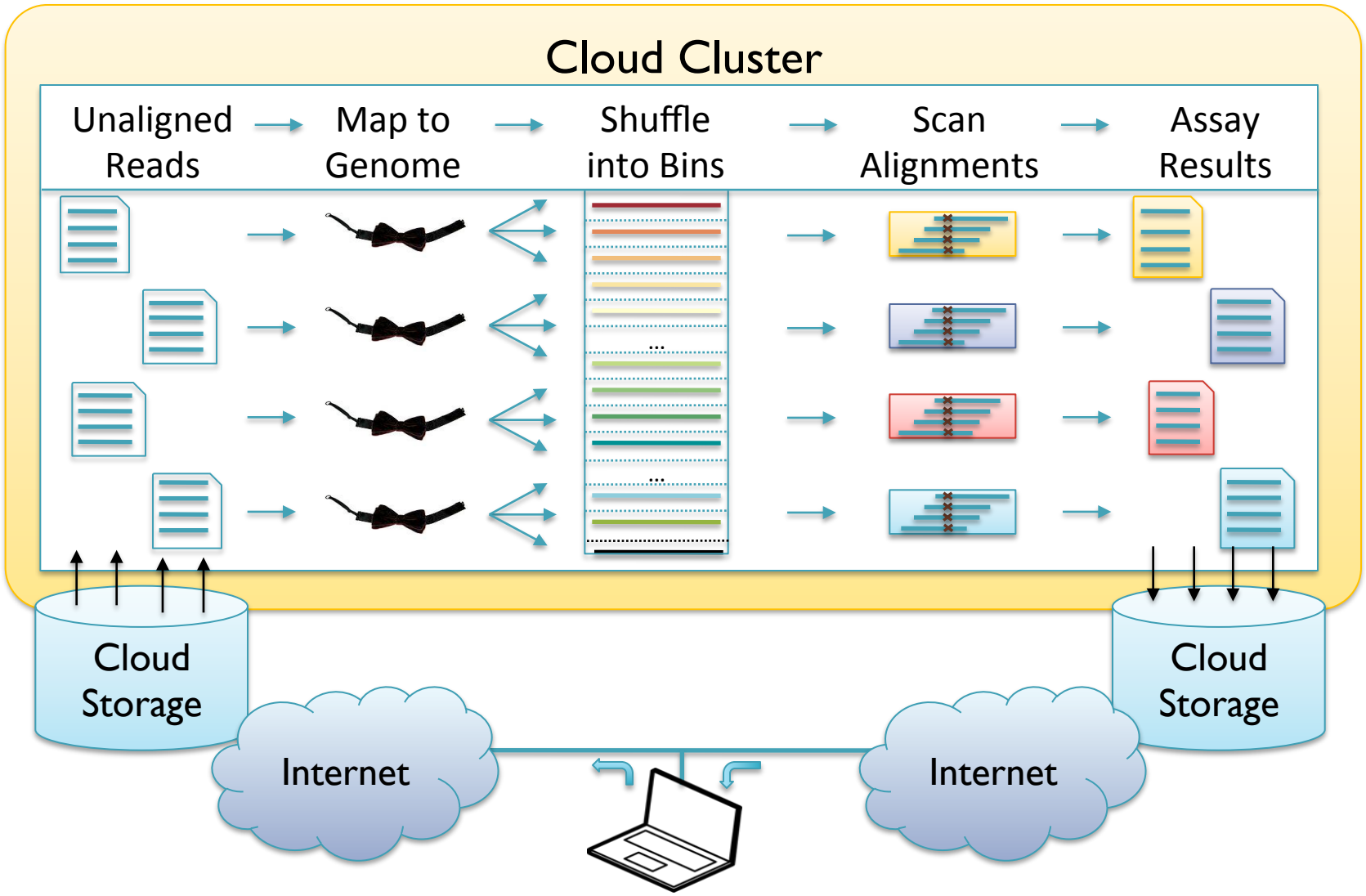
Construct the BWT of the human genome in 9 minutes

$GATTACA
A$GATTAC
ACA$GATT
ATTACA$G
CA$GATTA
GATTACA£
TACA$GAT
TTACA$GA

(Menon, Bhat, Schatz, 2011)

http://code.google.com/p/genome-indexing/

Map-Shuffle-Scan for Genomics



Cloud Computing and the DNA Data Race.
Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology.* **28**:691-693

Jnomics: Cloud-scale genomics

Matt Titmus*, Anirudh Aithal, James Gurtowski, Hayan Lee, Mitchell Bekritsky

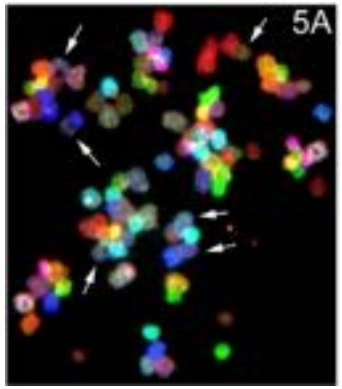
- Rapid execution of parallelized analysis pipelines
  - Mapping: BWA, Novoalign; SNPs/Indels: SAMTools; SV: Hydra
- Format agnostic: seamless read/write of common formats
  - BAM, SAM, BED, fastq, fasta represented in internal JAM format
  - Sorting, merging, filtering, selection, etc
- Open-source Java API for adding new components.
  - Existing pipelines: Quake, Crossbow, Myrna, Contrail
  - Mapping & Mappability: Bowtie, GMA
  - CNV: CNVnator, RDxplorer
  - Expression analysis: Tophat/Cufflinks, RSEQTools
  - …

Jnomics case study:
Structural variations in esophageal cancer

- Structural variations are common to many forms of cancer
  - Indels, Inversions, CNVs, Translocations of more than a single basepair
  - "An analysis of available data shows that gene fusions occur in all malignancies, and that they account for 20% of human cancer morbidity."
    - Mitelman *et al.* (2007) The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer.* 7:223-245

- Traditionally identified through cytogenetic imaging & microarrays
  - FISH, CGH, SOMA, etc

- Recent trend is to use sequencing to identify SVs
  - Decreased cost, improved resolution
  - Potential exists for basepair resolution of events

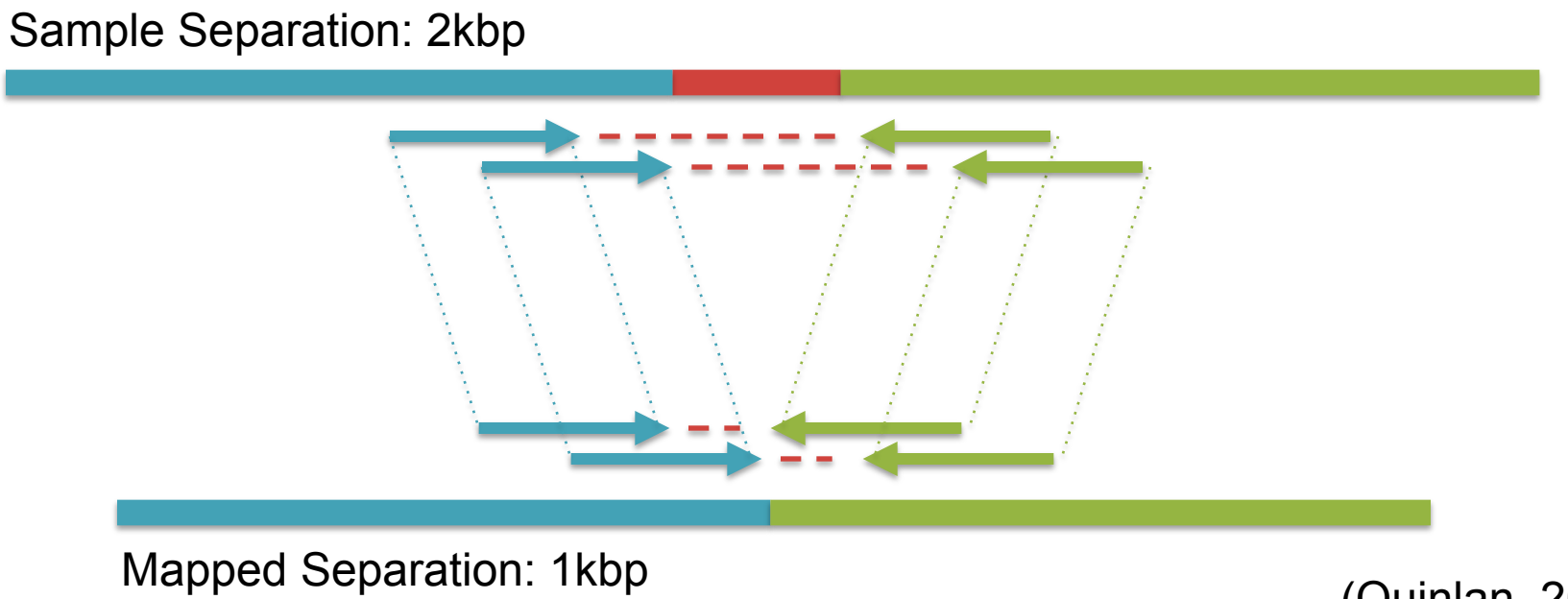


Applications of SKY in cancer cytogenetics
Bayani, JM, Squire, JA (2002) Cancer Invest. 20(3):373-86.

Hydra Discordant Pair Analysis

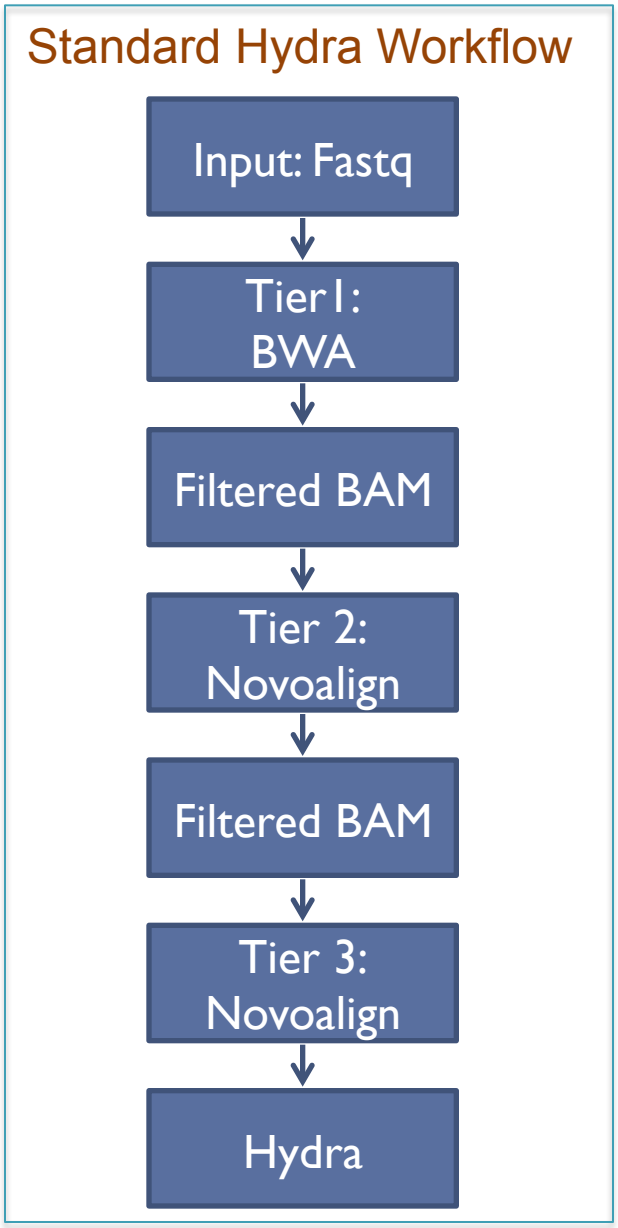
Illumina sequencing generates reads in pairs from both ends of a fragment with a known separation

1. Sequence diseased sample using paired-end/mate-pair protocol
2. Map reads from sample to reference genome
3. If a pair maps unexpectedly far away or with unexpected orientation, there is a SV between the reads
4. Cluster pairs to pinpoint breakpoints



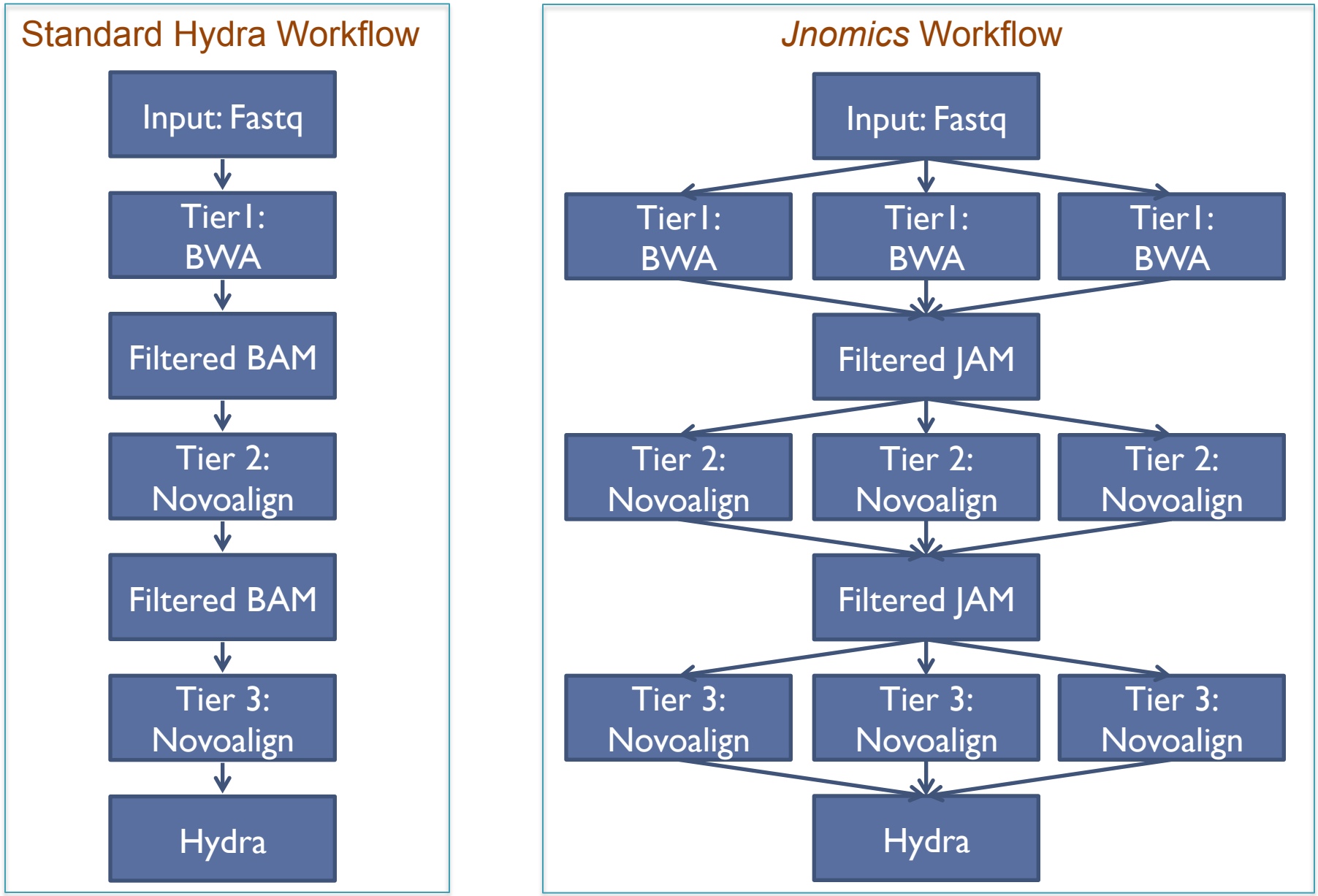
(Quinlan, 2010)

Jnomics SV Workflow



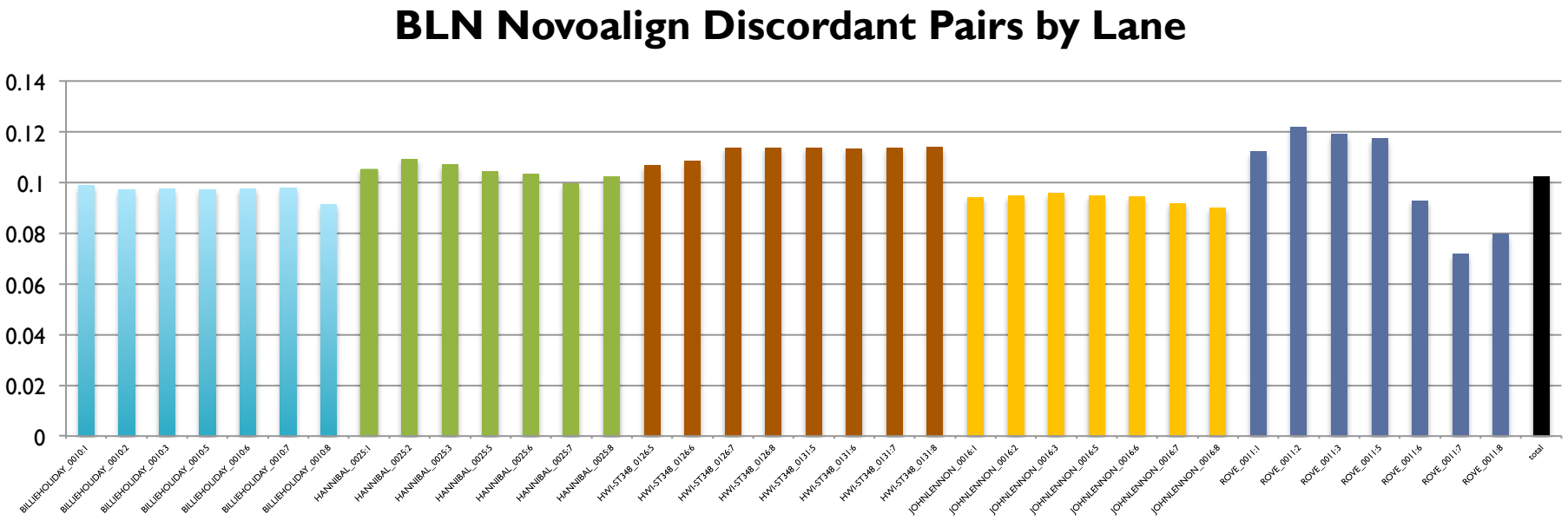
- Tiered alignment
  - Progressively increases sensitivity
  - After each stage select discordant pairs
- Cluster remaining discordant pairs
  - Require multiple pairs to filter random chimeric pairs
  - Apply coverage threshold to control sensitivity / specificity
- Overall workflow is resource intensive
  - Several weeks per single genome
  - Opportunities for parallelism for some stages, but batch computing is not sufficient

Jnomics SV Workflow



Pair Analysis of Esophageal Cancer

- BLN: Normal Tissue
  - 1.56B reads, 10% discordant pairs using Novoalign (16% after BWA)
- BLB: Barrett's Esophagus
  - 1.84B reads, 11% discordant pairs using Novoalign (17% after BWA)
- BLL: Invasive Adenocarcinoma
  - 1.77B reads, 14% discordant pairs using Novoalign (50% after BWA)



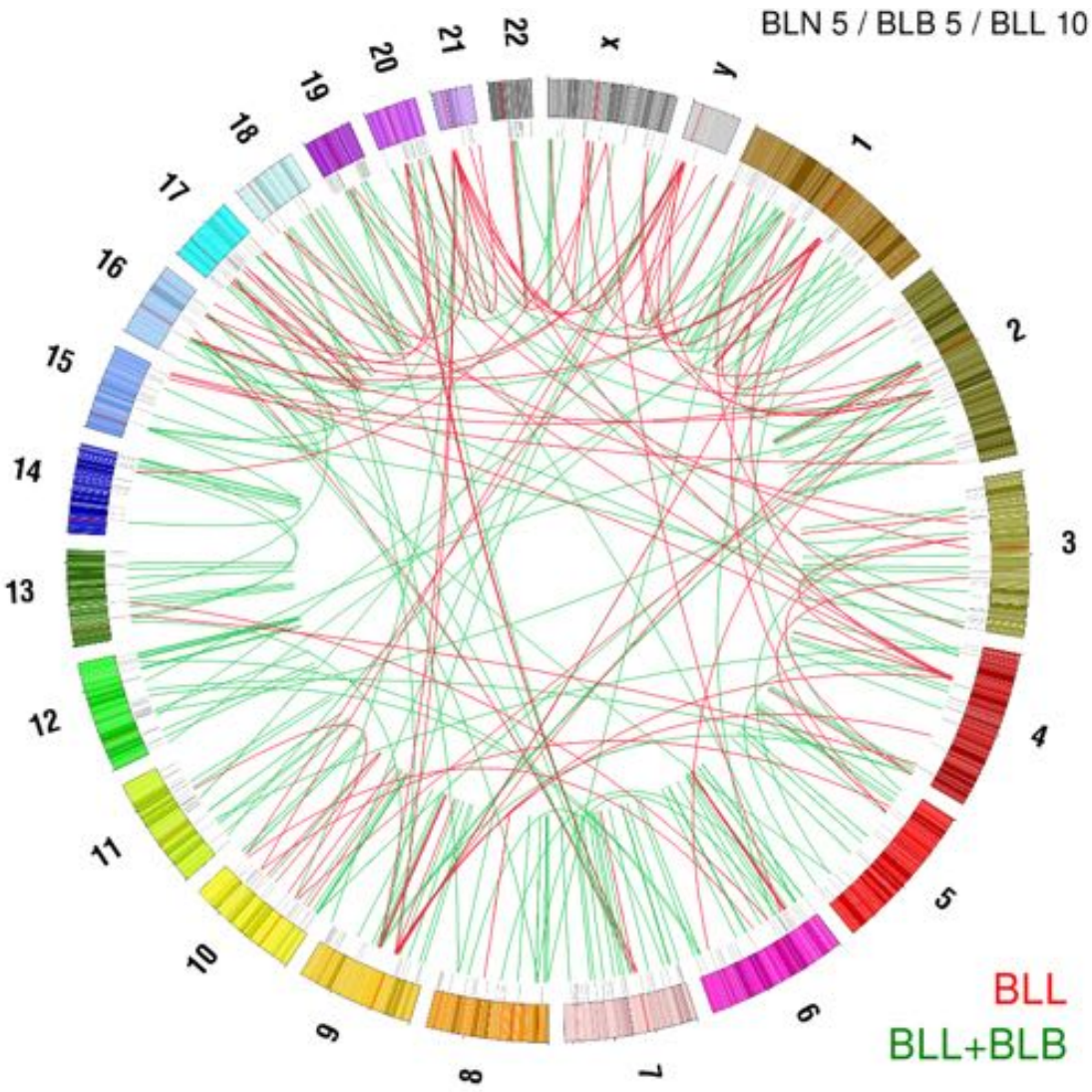
Jnomics Structural Variations

Circos plot of high confidence SVs specific to esophageal cancer sample

- Red: SVs specific to tumor
- Green: SVs in both diseased and tumor samples

Detailed analysis of disrupted genes and fusion genes in progress

- Preliminary analysis shows many promising hits to known cancer genes



Summary

- Staying afloat in the data deluge means computing in parallel
  - Hadoop + Cloud computing is an attractive platform for large scale sequence analysis and computation
- Significant obstacles ahead
  - Time and expertise required for development
  - Transfer, storage capabilities
  - Privacy / security requirements
- Emerging technologies are a great start, but we need continued research
  - Need integration across disciplines
  - A word of caution: new technologies are new

Acknowledgements

Schatzlab
Mitch Bekritsky
Matt Titmus
Hayan Lee
James Gurtowski
Anirudh Aithal
Rohith Menon
Goutham Bhat

CSHL
Dick McCombie
Melissa Kramer
Eric Antonio
Mike Wigler
Zach Lippman
Doreen Ware
Ivan Iossifov

JHU
Steven Salzberg
Ben Langmead
Jeff Leek

NBACC
Adam Phillipy
Sergey Koren

Univ. of Maryland
Mihai Pop
Art Delcher
Jimmy Lin
David Kelley
Dan Sommer
Cole Trapnell

(GIGA)^n SCIENCE

Now taking submissions!

Large-Scale Data Journal/Database

GigaScience aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish 'big-data' studies from the entire spectrum of life and biomedical sciences. A novel publication format links standard manuscript publication with an extensive database that hosts all associated data, provides data analysis tools, cloud-computing resources, and gives all datasets a DOI as a citable and trackable data publication mark.

For more information: editorial@gigasciencejournal.com @gigascience

华大基因 BGI

www.gigasciencejournal.com

BioMed Central
The Open Access Publisher

Thank You!

http://schatzlab.cshl.edu
@mike_schatz / #btg11