

Entering the era of mega-genomics

Michael Schatz

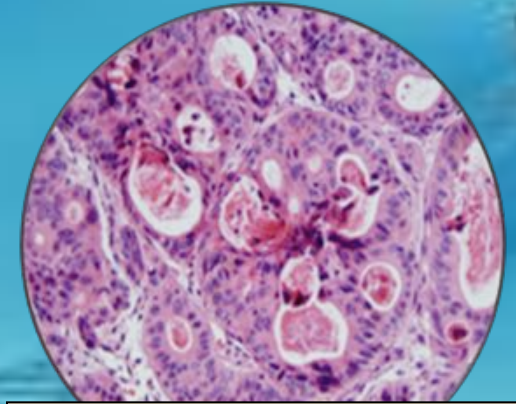
March 2, 2012
UNC Charlotte



Schatz Lab Overview



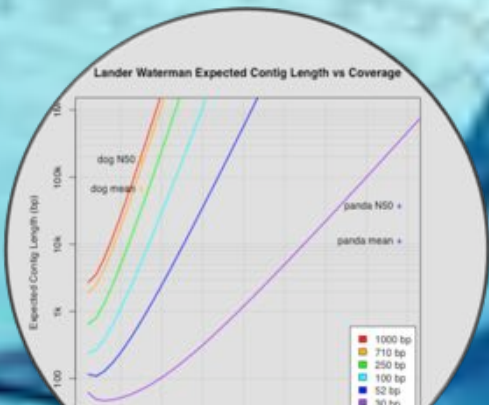
Computation



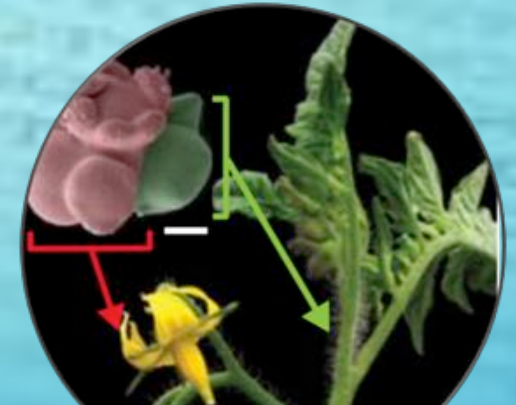
Human Genetics



Sequencing



Modeling



Plant Genomics

Outline



1. Milestones in genomics

1. Sanger to nanopore
2. 21st Century Mega-Genomics

2. Applications of mega-genomics

1. Single molecule sequencing & assembly
2. Cloud-scale resequencing
3. De novo mutations in autism

Milestones in Genomics: Zeroth Generation Sequencing

Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage Φ X174 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

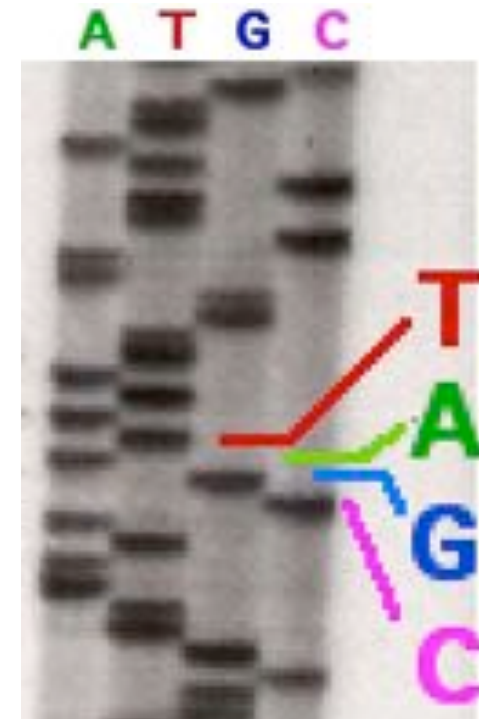
A DNA sequence for the genome of bacteriophage Φ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including 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 Φ X174 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 techniques¹⁻³, is A-B-C-D-E-F-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 Φ X 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 protein⁴ (positions 2,362-2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed^{5,6} and Schott⁷ synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intergenic region between the F and G genes, using DNA polymerase and ³²P-labelled triphosphates⁸. The ribo-substitution technique⁹ facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method¹⁰. Suitable synthetic primers are, however, difficult to prepare and an

1977
1st Complete Organism
Bacteriophage ϕ X174
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 ϕ X174 DNA
Sanger et al. (1977) Nature. 265: 687 - 695

Milestones in Genomics: First Generation Sequencing



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

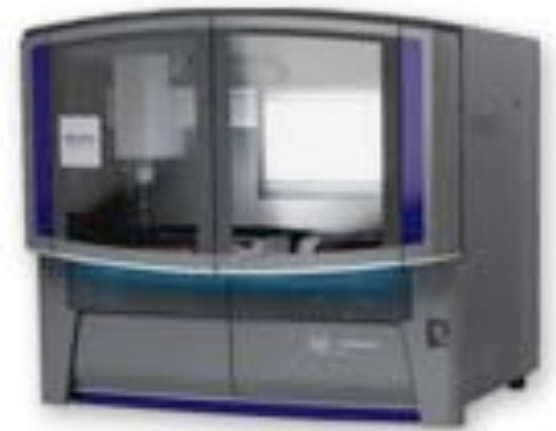
Milestones in Genomics: Second Generation Sequencing



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



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



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

Milestones in Genomics: Third Generation Sequencing



2010

Ion Torrent

Postlight Sequencing

Current Specs (Ion 318):

11M 300bp reads / run =

>1Gbp / day



2011

Pacific Biosciences

SMRT Sequencing

Current Specs (RS):

50k 2kbp reads / run =

>200Mbp / day



2012

Oxford Nanopore

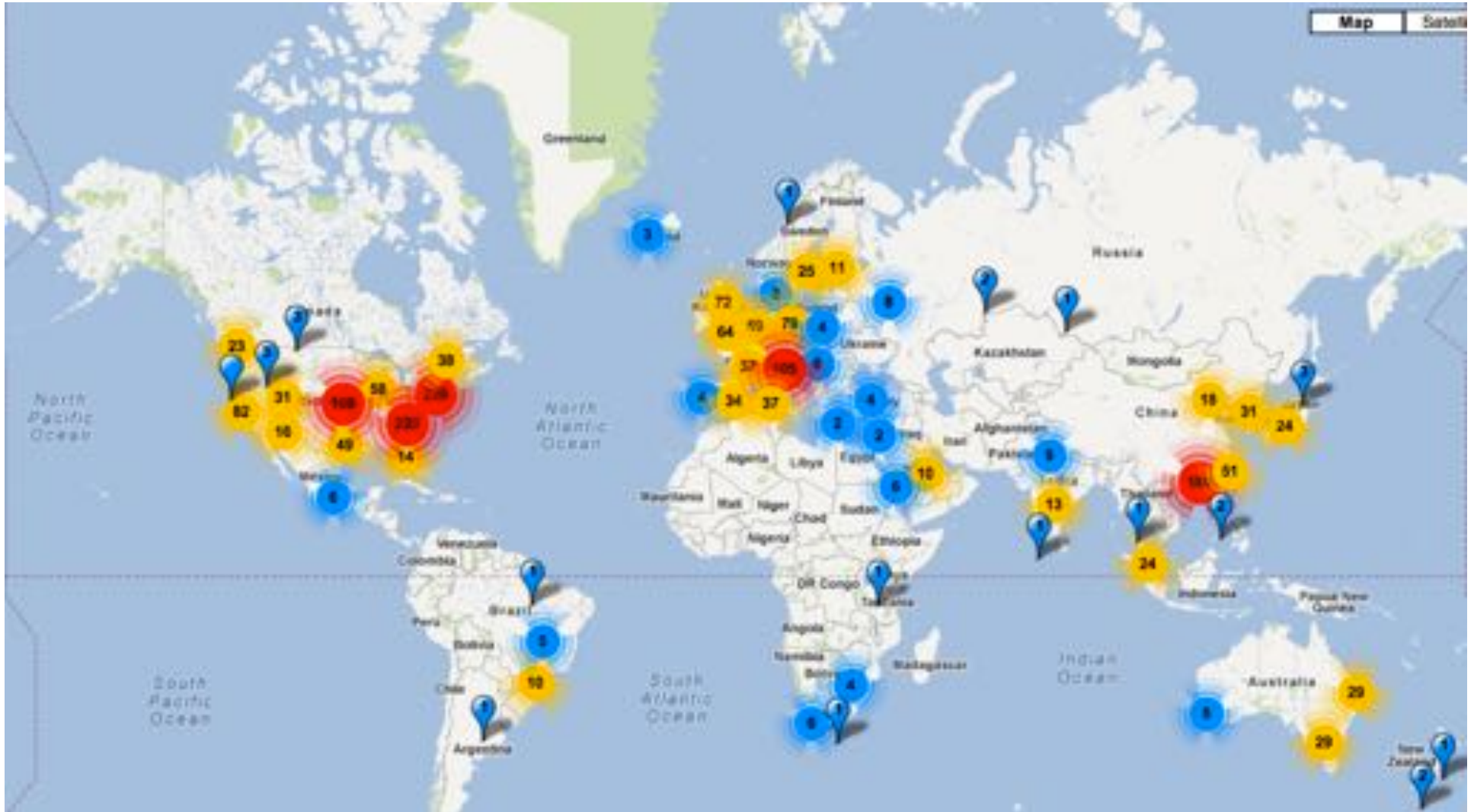
Nanopore sensing

Current Specs (GridIron):

Reads up to 48kbp

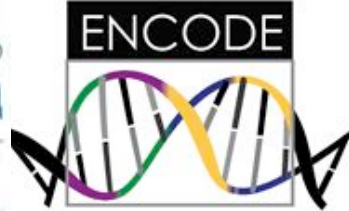
Many GB / day

Sequencing Centers

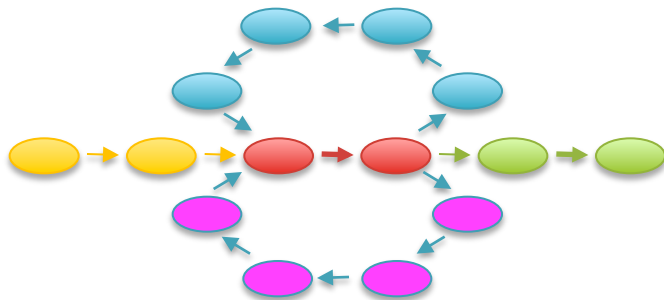


Next Generation Genomics: World Map of High-throughput Sequencers
<http://pathogenomics.bham.ac.uk/hts/>

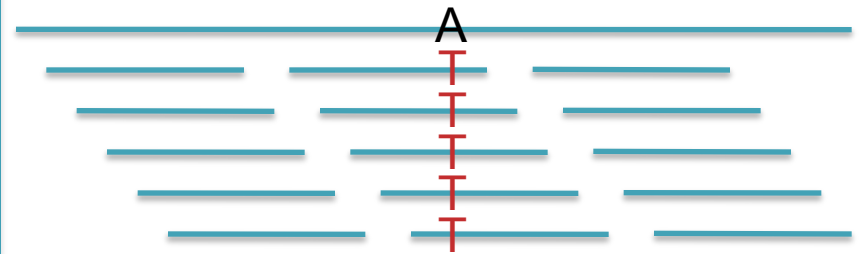
The rise of mega-genomics



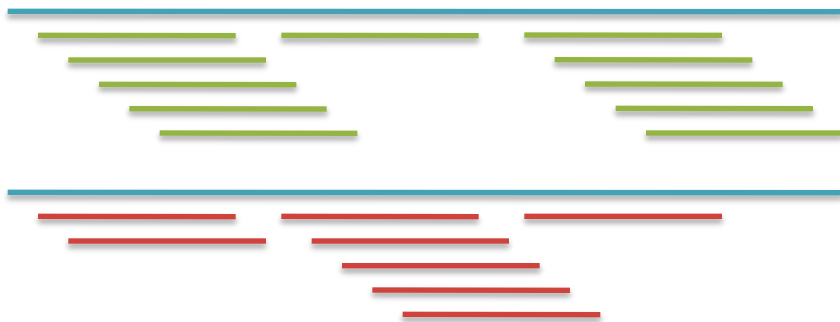
De novo Assembly



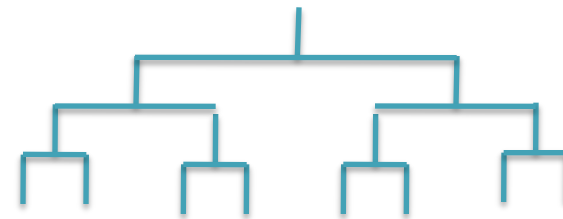
Alignment & Variations



Differential Analysis



Phylogeny & Evolution



Mega-Genomics Challenges



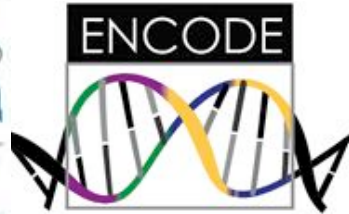
The foundations of genomics will continue to be *observation, experimentation, and interpretation*

- Technology will continue to push the frontier
- Measurements will be made *digitally* over large populations, at extremely high resolution, and for diverse applications

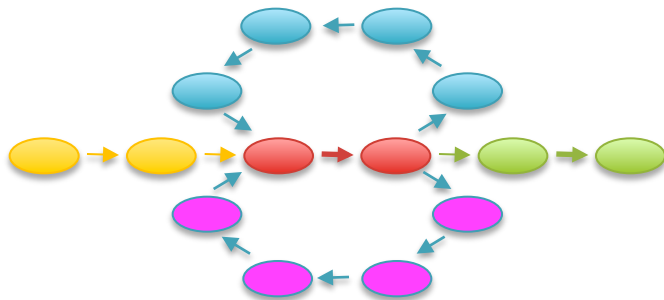
Rise in Quantitative Demands

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

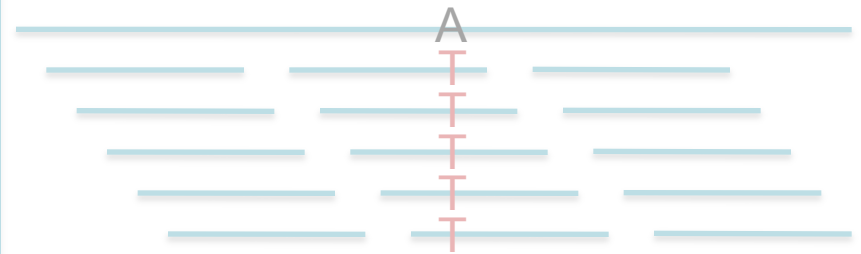
The rise of mega-genomics



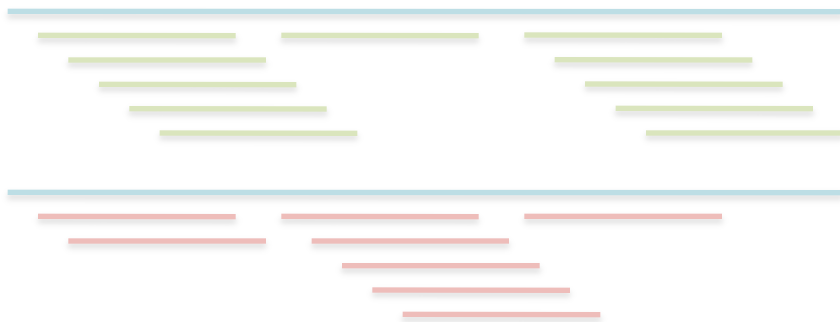
De novo Assembly



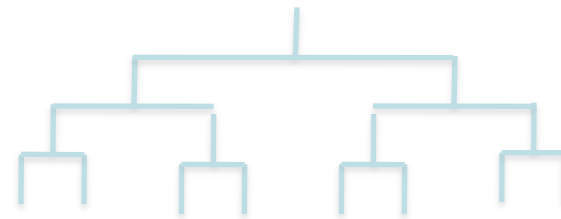
Alignment & Variations



Differential Analysis

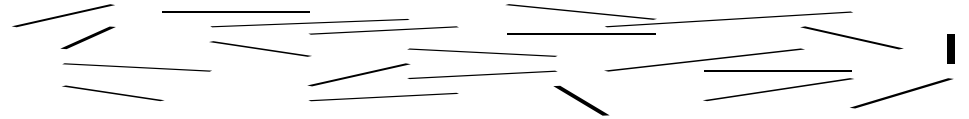


Phylogeny & Evolution



Assembling a Genome

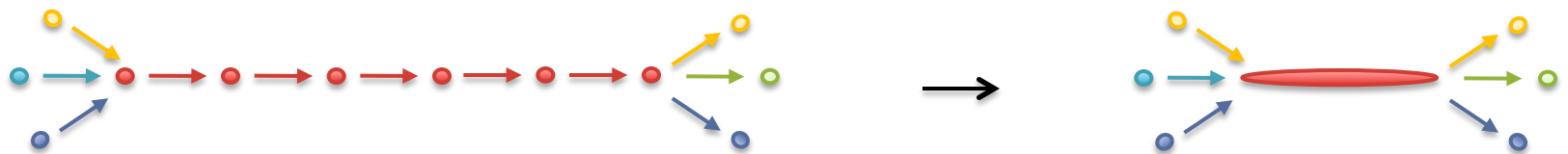
1. Shear & Sequence DNA



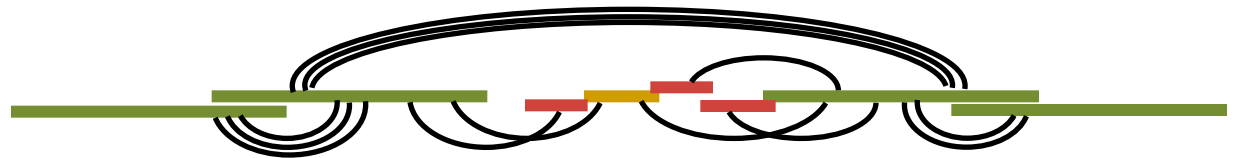
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTTCGCATATCCGGT...

3. Simplify assembly graph

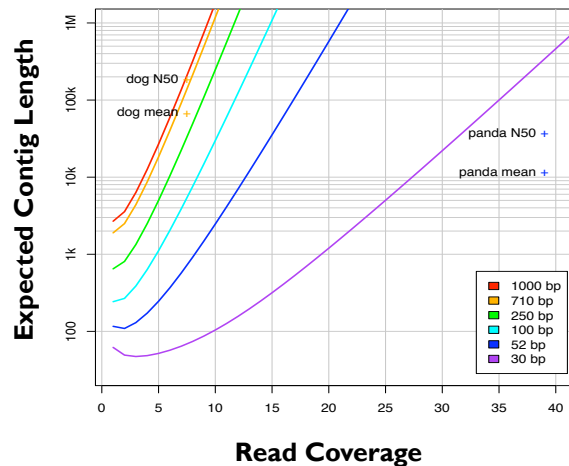


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

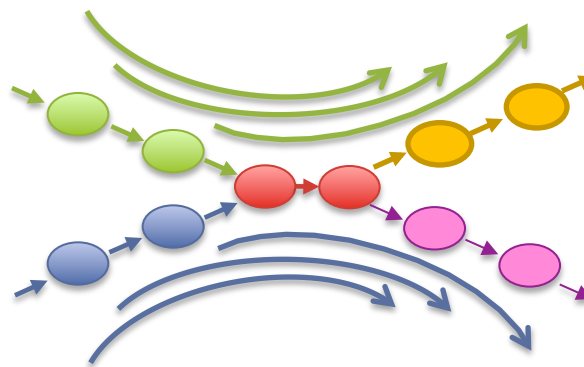
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

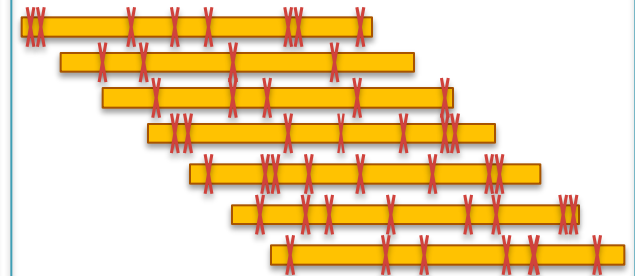
Read Length



Reads & mates must be longer than the repeats

- Short reads will have **false overlaps** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Assembly of Large Genomes using Second Generation Sequencing

Schatz MC, Delcher AL, Salzberg SL (2010) *Genome Research*. 20:1165-1173.

Hybrid Sequencing



Illumina

Sequencing by Synthesis

High throughput (60Gbp/day)

High accuracy (~99%)

Short reads (~100bp)



Pacific Biosciences

SMRT Sequencing

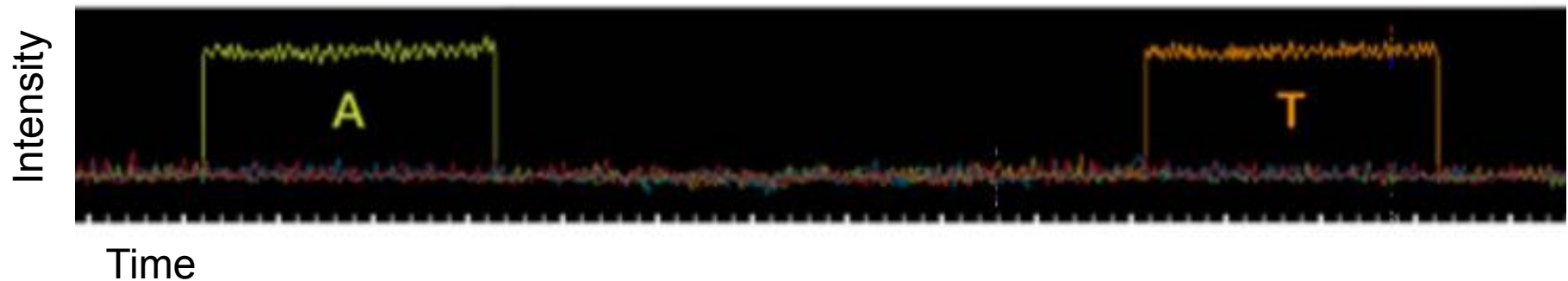
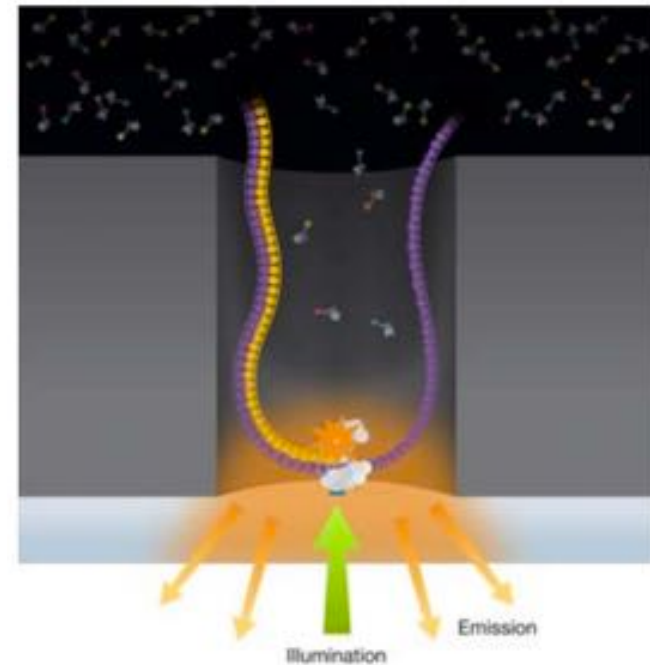
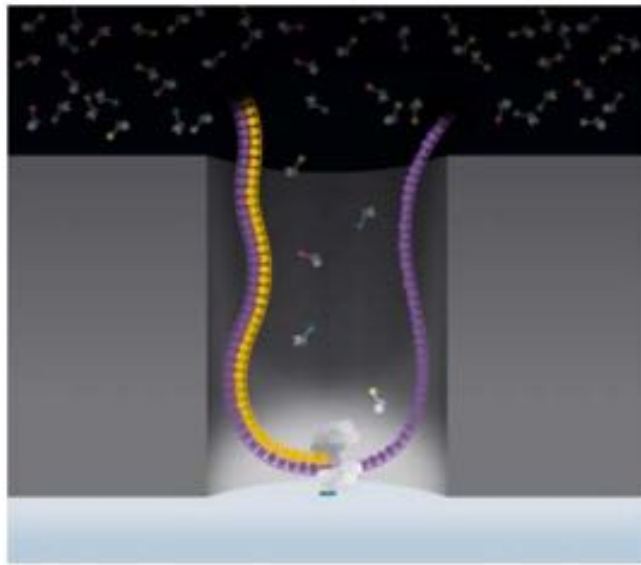
Lower throughput (600Mbp/day)

Lower accuracy (~85%)

Long reads (1-2kbp+)

SMRT Sequencing

Imaging of fluorescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



PacBio Error Correction

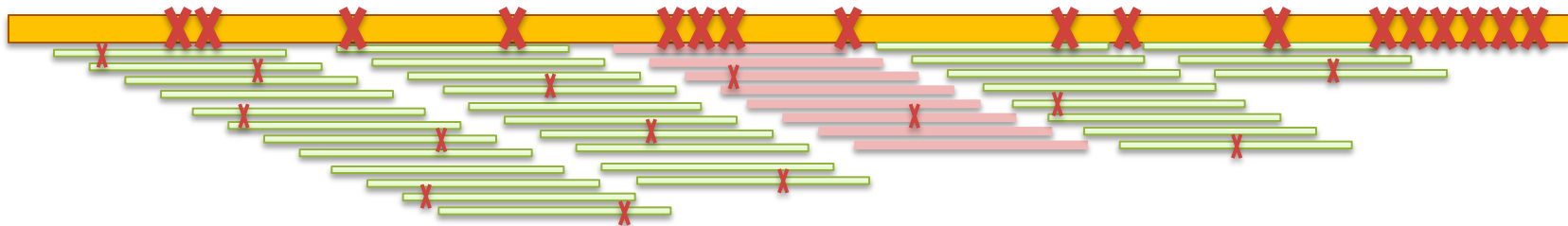
<http://wgs-assembler.sf.net>



I. Correction Pipeline

1. Map short reads (SR) to long reads (LR)
2. Trim LR at coverage gaps
3. Compute consensus for each LR

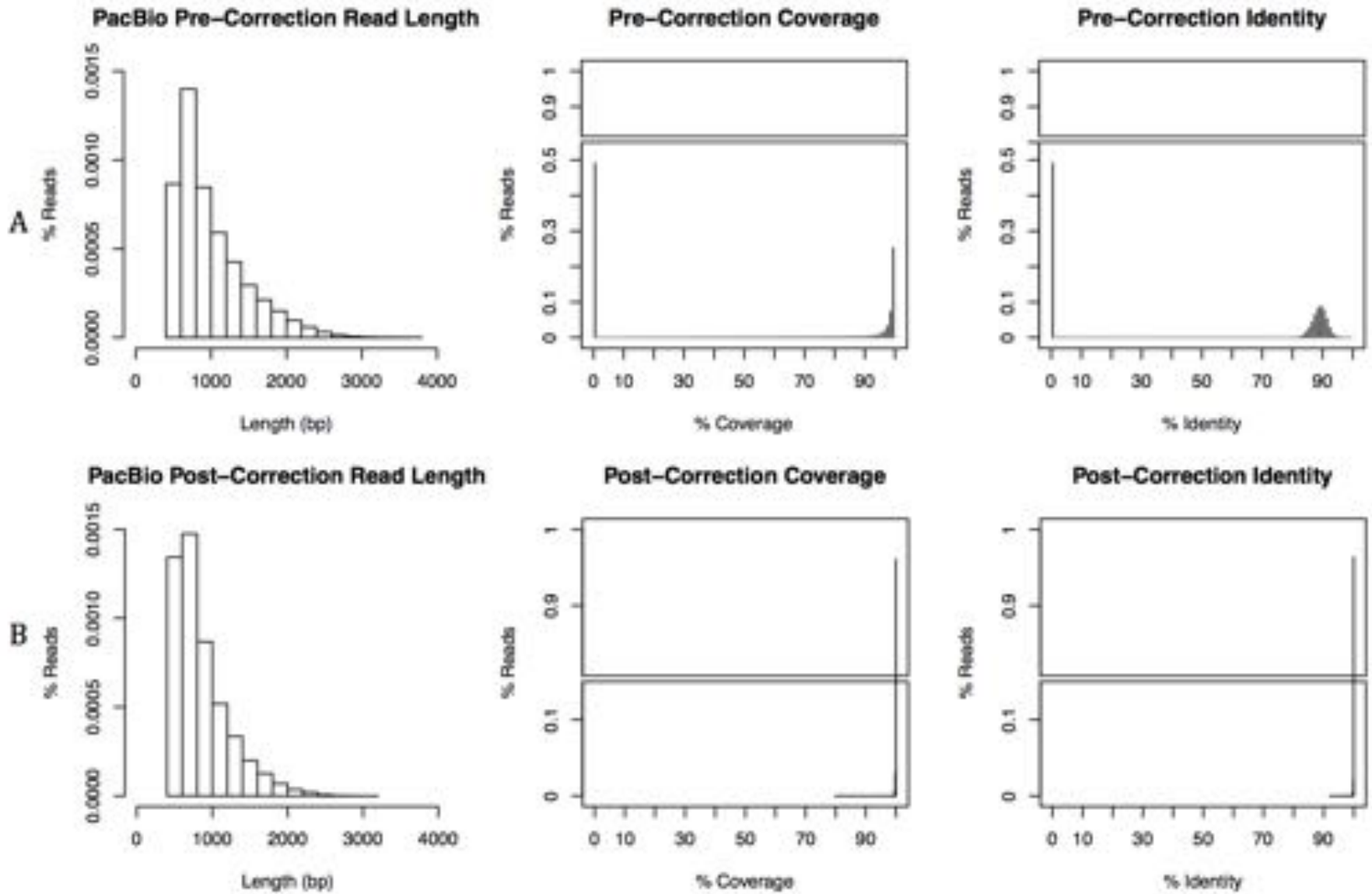
2. Error corrected reads can be easily assembled, aligned



Hybrid error correction and de novo assembly of single-molecule sequencing reads.

Koren, S, Schatz, MC, Walenz, BP, Martin, J, Howard, J, Ganapathy, G, Wang, Z, Rasko, DA, McCombie, VWR, Jarvis, ED, Phillippy, AM. (2012) *Under Review*

Error Correction Results

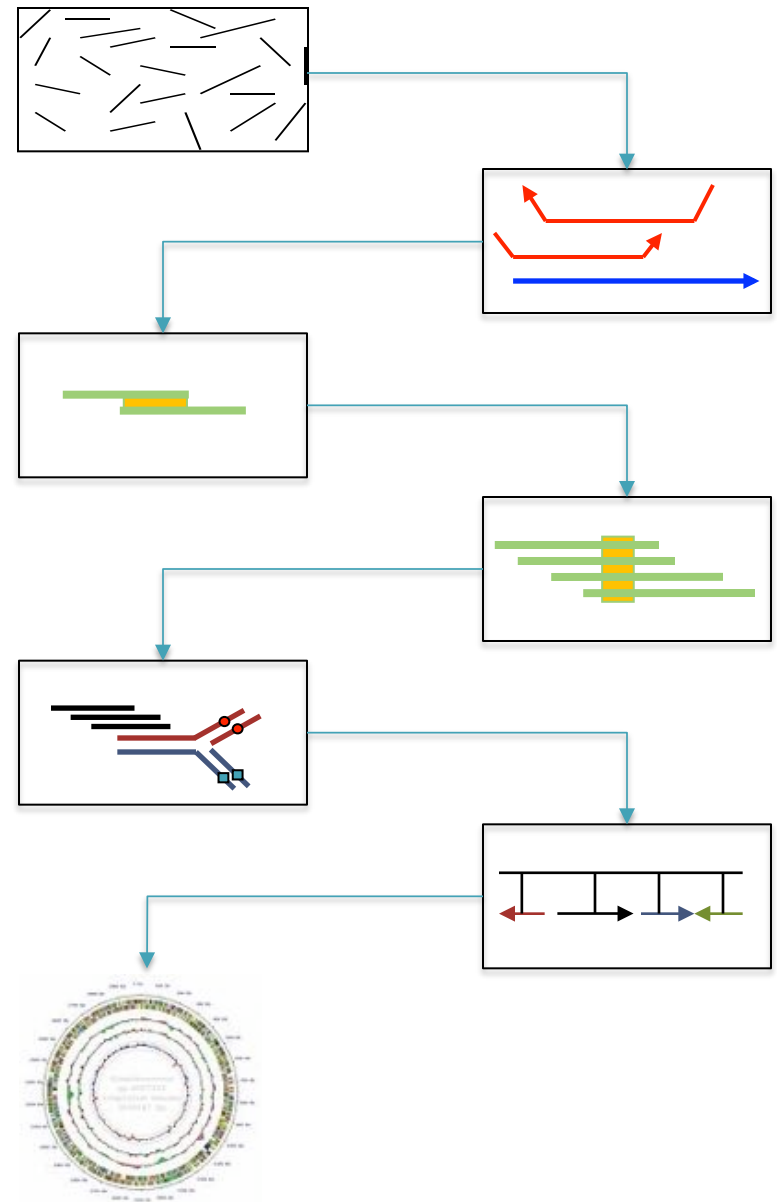


Correction results of 20x PacBio coverage of *E. coli* K12 corrected using 50x Illumina

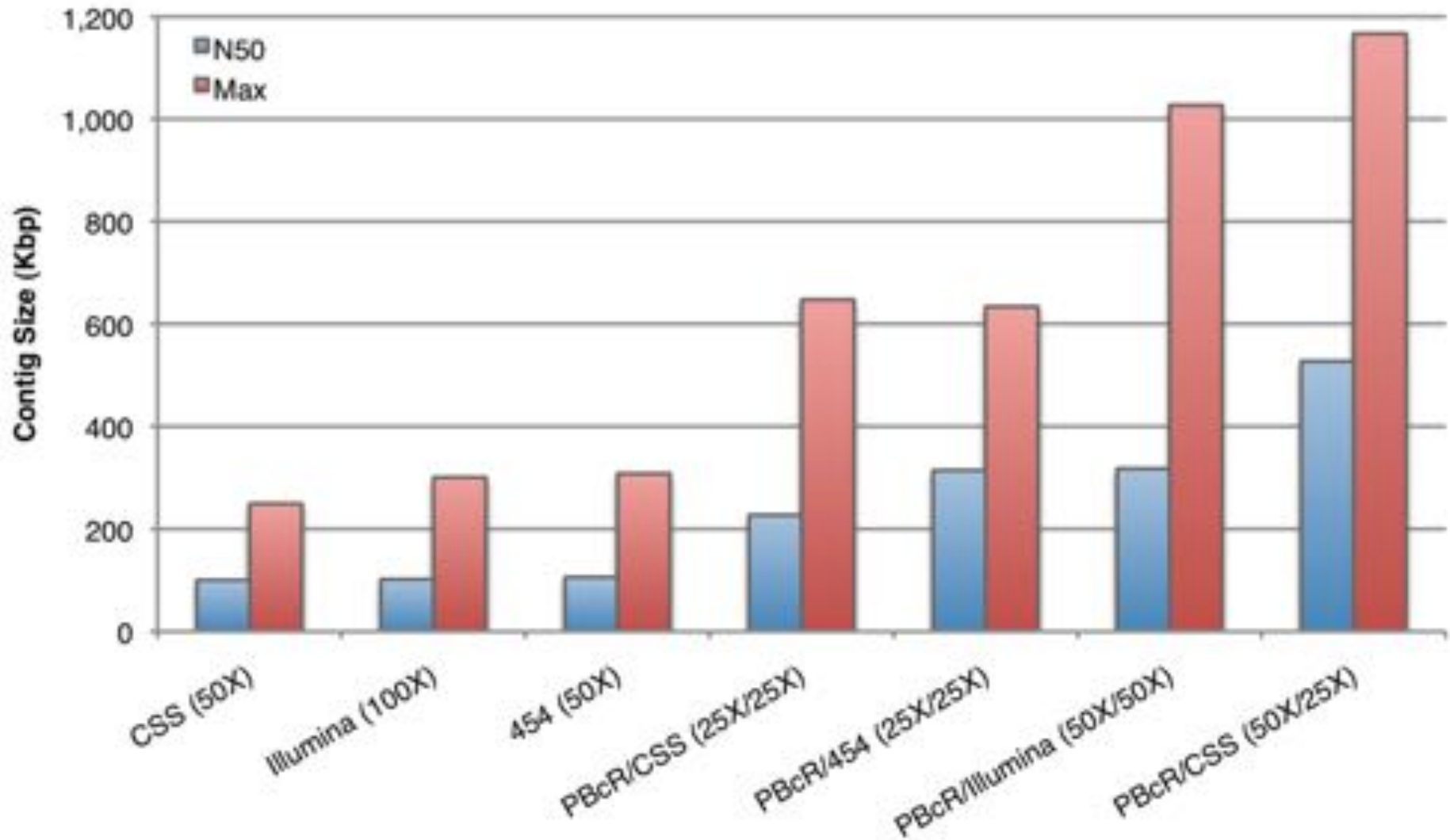
Celera Assembler

<http://wgs-assembler.sf.net>

1. Pre-overlap
 - Consistency checks
2. Trimming
 - Quality trimming & partial overlaps
3. Compute Overlaps
 - Find high quality overlaps
4. Error Correction
 - Evaluate difference in context of overlapping reads
5. Unitigging
 - Merge consistent reads
6. Scaffolding
 - Bundle mates, Order & Orient
7. Finalize Data
 - Build final consensus sequences



Assembly Results



SMRT-assembly results of 50x PacBio corrected coverage of E. coli K12
Long reads lead to **contigs** over 1Mbp

SMRT-Assembly Results



Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
<i>Lambda</i> NEB3011 (median: 727 max: 3 280)	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
<i>E. coli</i> K12 (median: 747 max: 3 068)	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.36%) *
	PacBio PBcR 18X		4 465 533	77	239 058 / 238 224	71 479 / 68 309 (95.57%) *
	Both 18X PacBio PBcR + Illumina 50X 500bp		4 576 046	65	238 272 / 238 224	93 048 / 89 431 (96.11%) *
<i>E. coli</i> C227-11 (median: 1 217 max: 14 901)	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
	PacBio 25X PBcR (corrected by 25X CCS)		5 207 946	80	357 234	98 774
	Both PacBio PBcR 25X + CCS 25X		5 269 158	39	647 362	227 302
	PacBio 50X PBcR (corrected by 50X CCS)		5 445 466	35	1 076 027	376 443
	Both PacBio PBcR 50X + CCS 25X		5 453 458	33	1 167 060	527 198
	Manually Corrected ALLORA Assembly ⁸		5 452 251	23	653 382	402 041
<i>S. cerevisiae</i> S228c (median: 674 max: 5 994)	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
	PacBio PBcR 13X		11 110 420	224	224 478 / 217 704	62 898 / 54 633 (86.86%) *
	Both PacBio PBcR 13X + Illumina 50X 300bp		11 286 932	177	262 846 / 260 794	82 543 / 59 792 (72.44%) *
<i>Melospiza ardensis</i> (median 997, max 13 079)	Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)	1.23 Gbp	1 023 532 850	24 181	1 050 202	47 383
	454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)		999 168 029	16 574	751 729	75 178
	454 15.4X + PacBio PBcR 3.75X		1 071 356 415	15 081	1 238 843	99 573

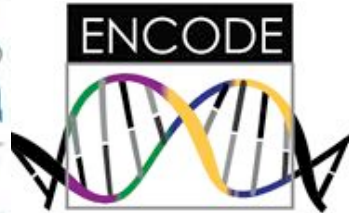
Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case

Transcript Alignment

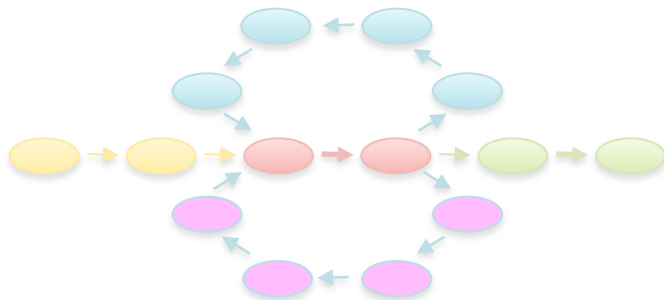


- Long-read single-molecule sequencing has potential to directly sequence full length transcripts
 - Raw reads and raw alignments (red) have many spurious indels inducing false frameshifts and other artifacts
 - Error corrected reads almost perfectly match the genome, pinpointing splice sites, identifying alternative splicing
- New collaboration with Gingeras Lab looking at splicing in human

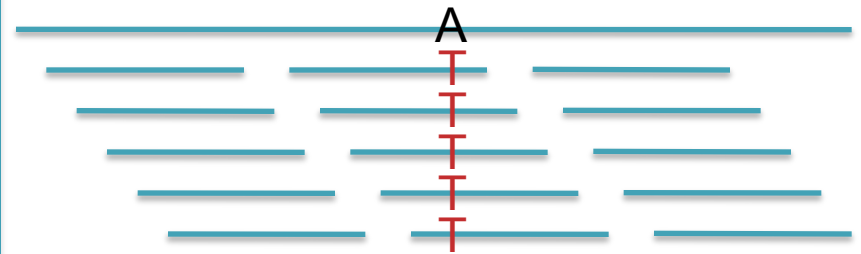
The rise of mega-genomics



De novo Assembly



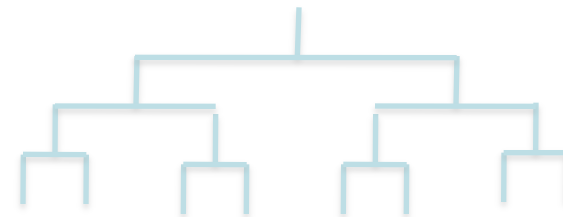
Alignment & Variations



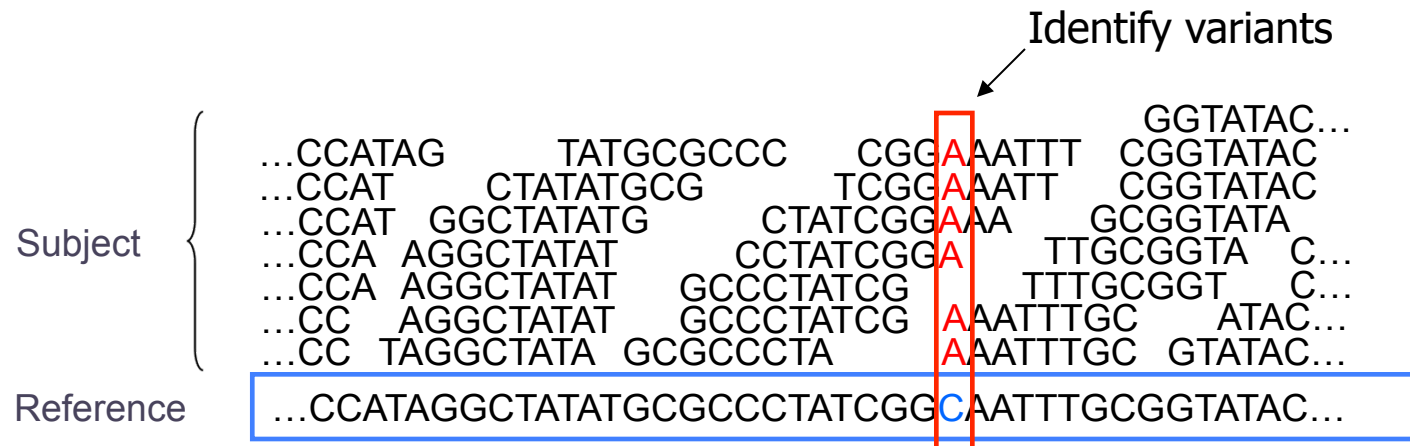
Differential Analysis



Phylogeny & Evolution



Short Read Mapping

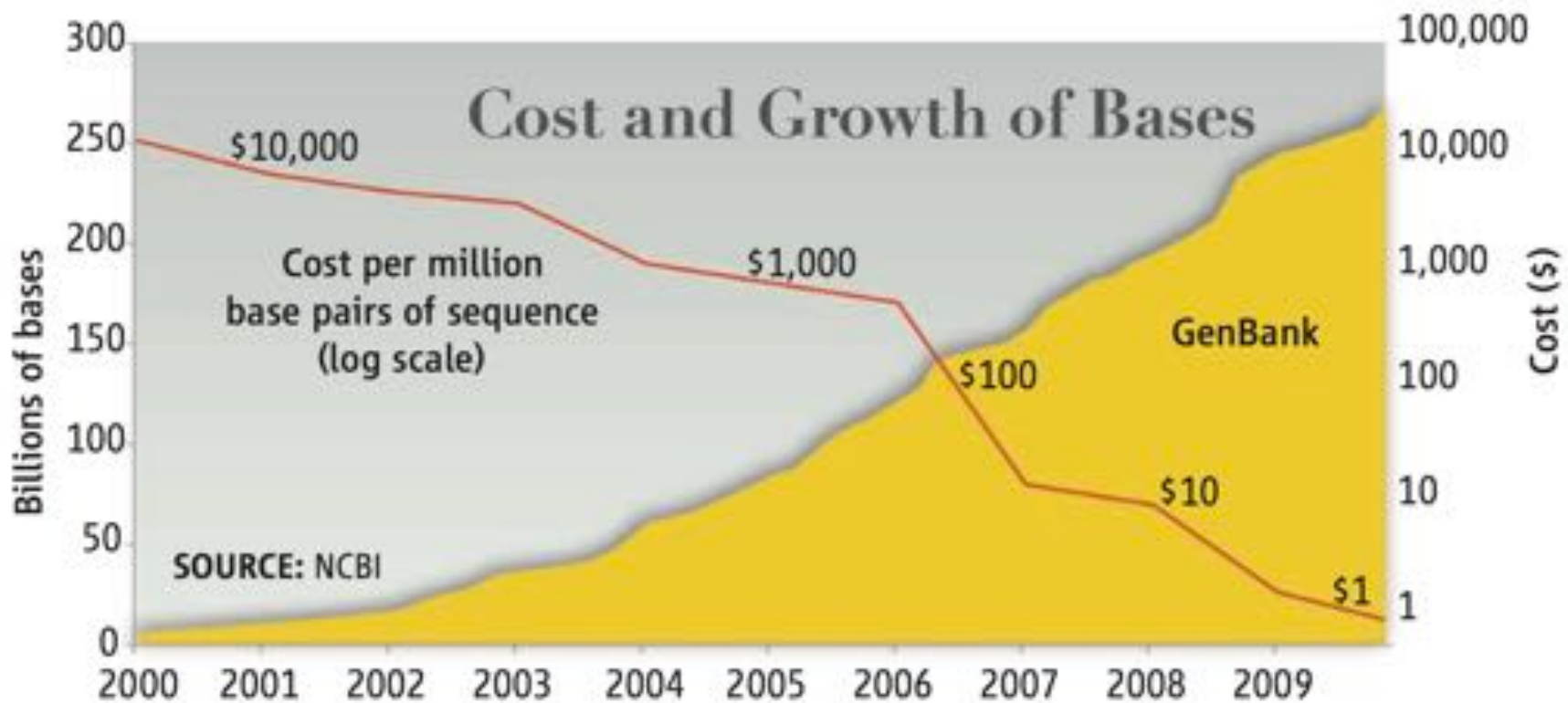


- Given a reference and many subject reads, report one or more “good” end-to-end alignments per alignable read
 - Find where the read most likely originated
 - Fundamental computation for many assays
 - Genotyping RNA-Seq Methyl-Seq
 - Structural Variations Chip-Seq Hi-C-Seq

- Desperate need for scalable solutions
 - Single human requires >1,000 CPU hours / genome

DNA Data Tsunami

*Current world-wide sequencing capacity exceeds 14Pbp/year
and is growing at 5x per year!*



"Will Computers Crash Genomics?"

Elizabeth Pennisi (2011) *Science*. 331(6018): 666-668.

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)
 - 946PB 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



CloudBurst

Highly Sensitive Short Read Mapping with MapReduce

100x speedup mapping on 96 cores @ Amazon

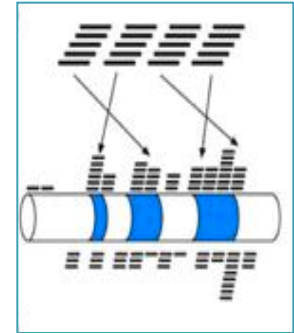
<http://cloudburst-bio.sf.net>

(Schatz, 2009)

Myrna

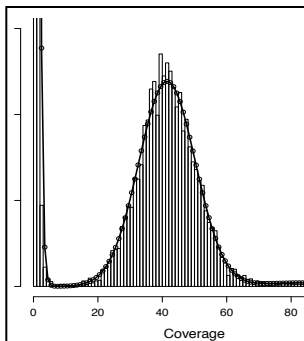
Cloud-scale differential gene expression for RNA-seq

Expression of 1.1 billion RNA-Seq reads in ~2 hours for ~\$66



(Langmead, Hansen, Leek, 2010)

<http://bowtie-bio.sf.net/myrna/>



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

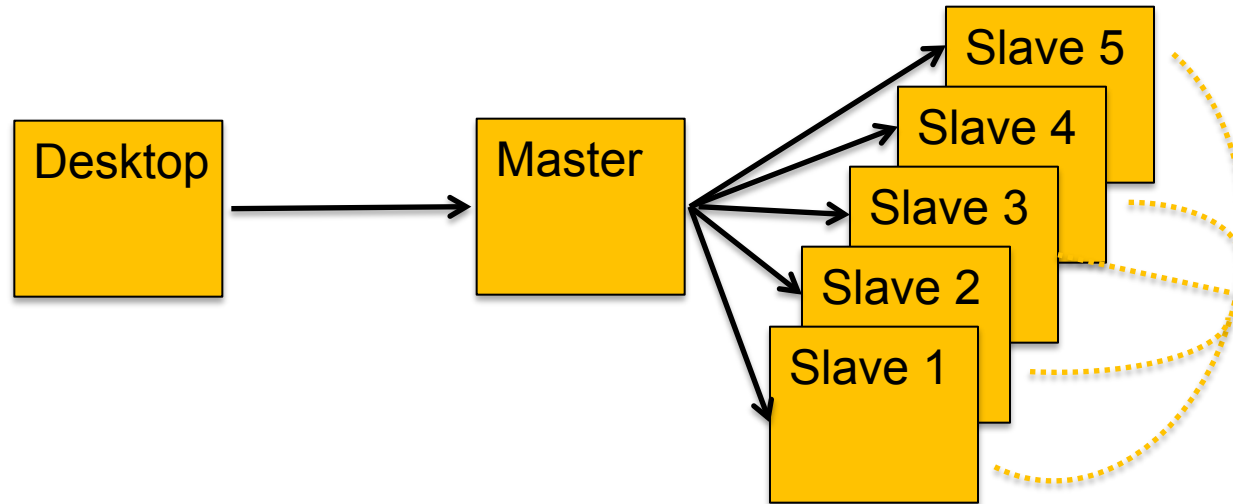
Construct the BWT of the human genome in 9 minutes

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

(Menon, Bhat, Schatz, 2011*)

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

System Architecture



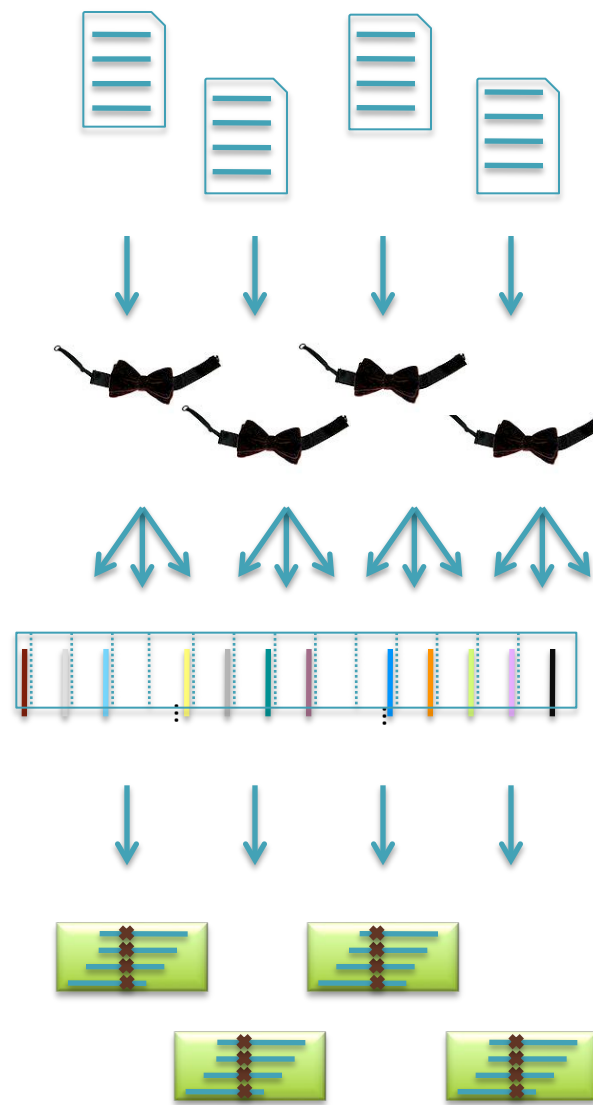
- Hadoop Distributed File System (HDFS)
 - Data files partitioned into large chunks (64MB), replicated on multiple nodes
 - Computation moves to the data, rack-aware scheduling
- Hadoop MapReduce system won the 2009 GreySort Challenge
 - Sorted 100 TB in 173 min (578 GB/min) using 3452 nodes and 4x3452 disks



Crossbow

<http://bowtie-bio.sourceforge.net/crossbow>

- Align billions of reads and find SNPs
 - Reuse software components: Hadoop Streaming
- Map: Bowtie (Langmead *et al.*, 2009)
 - Find best alignment for each read
 - Emit (chromosome region, alignment)
- Shuffle: Hadoop
 - Group and sort alignments by region
- Reduce: SOAPsnp (Li *et al.*, 2009)
 - Scan alignments for divergent columns
 - Accounts for sequencing error, known SNPs



Performance in Amazon EC2

<http://bowtie-bio.sourceforge.net/crossbow>

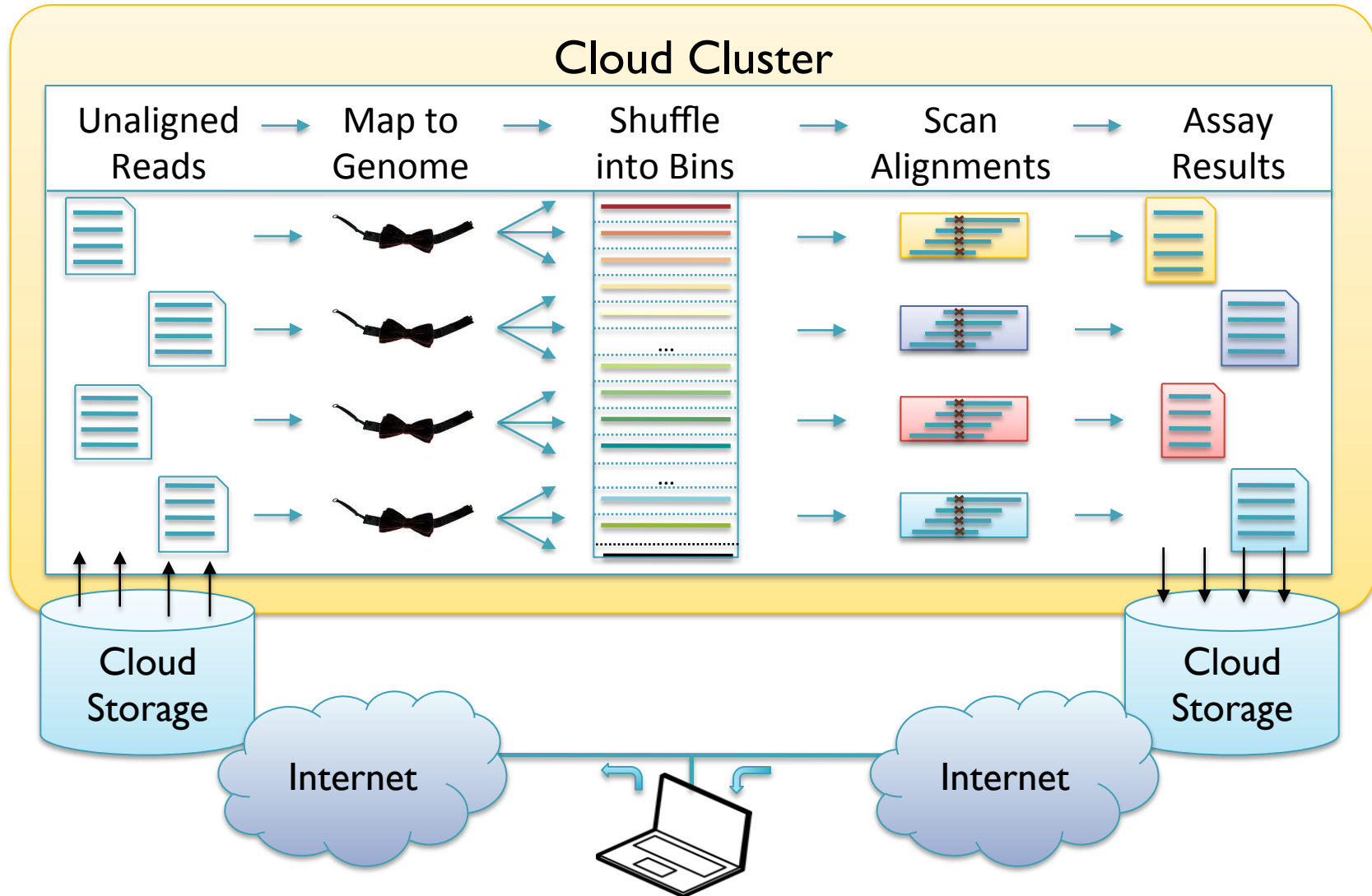
	Asian Individual Genome		
Data Loading	3.3 B reads	106.5 GB	\$10.65
Data Transfer	1h :15m	40 cores	\$3.40
Setup	0h : 15m	320 cores	\$13.94
Alignment	1h : 30m	320 cores	\$41.82
Variant Calling	1h : 00m	320 cores	\$27.88
End-to-end	4h : 00m		\$97.69

Discovered 3.7M SNPs in one human genome for ~\$100 in an afternoon.
Accuracy validated at >99%

Searching for SNPs with Cloud Computing.

Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology*. **10**:R134

Map-Shuffle-Scan for Genomics

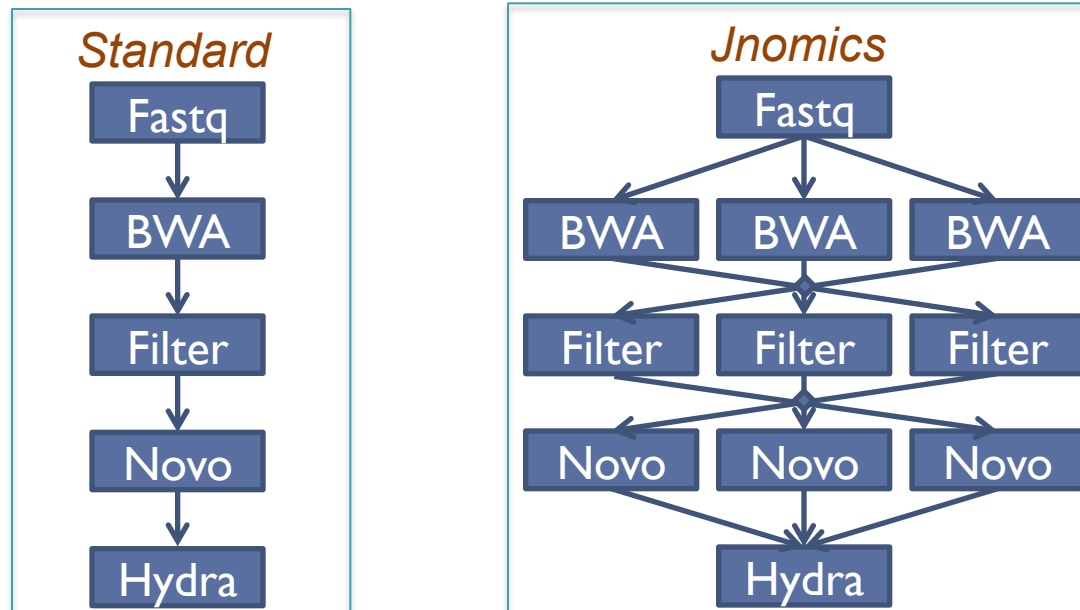


Cloud Computing and the DNA Data Race.

Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology*. **28**:691-693

Jnomics: Cloud-scale genomics

Matt Titmus, James Gurtowski, Michael Schatz



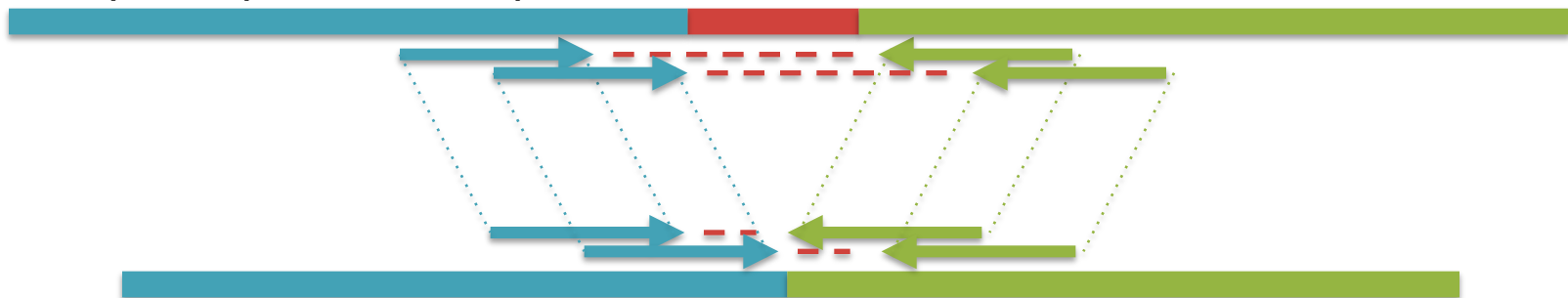
- Rapid parallel execution of NGS analysis pipelines
 - FASTX, BWA, Bowtie, Novoalign, SAMTools, Hydra
 - Sorting, merging, filtering, selection, of BAM, SAM, BED, fastq
 - Population analysis: Clustering, GWAS, Trait Inference

Answering the demands of digital genomics

Titmus, M.A., Schatz, M.C.. (2012) *Under Review*

Jnomics Structural Variations

Sample Separation: 2kbp



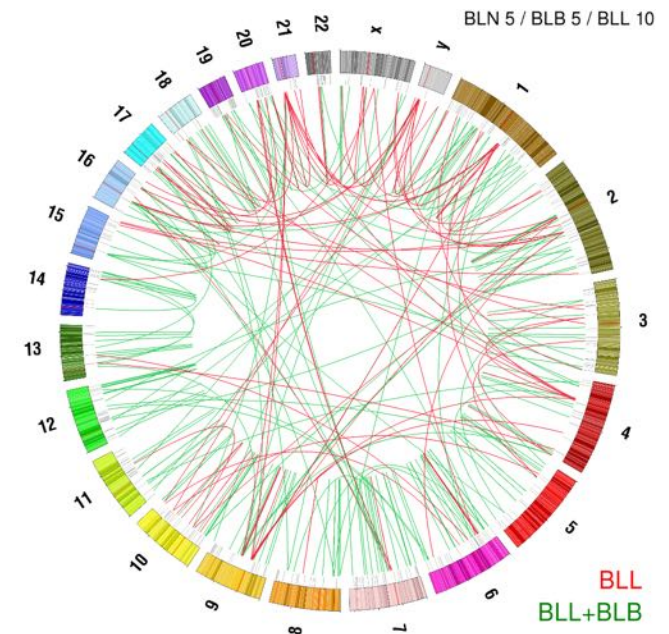
Mapped Separation: 1kbp

Discordant Pair Analysis

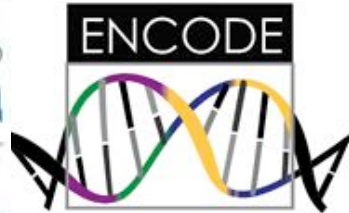
- Identify clusters of pairs too close or too far away indicating a SV

Circos plot of high confidence SVs specific to esophageal cancer sample

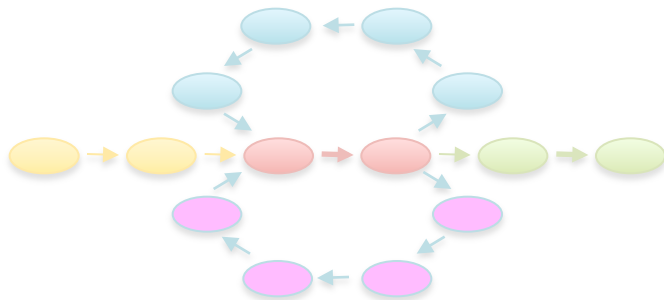
- Red: SVs specific to tumor
- Green: SVs in both diseased and tumor samples



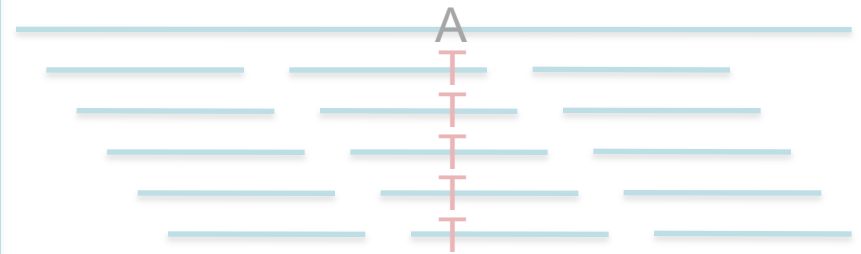
The rise of mega-genomics



De novo Assembly



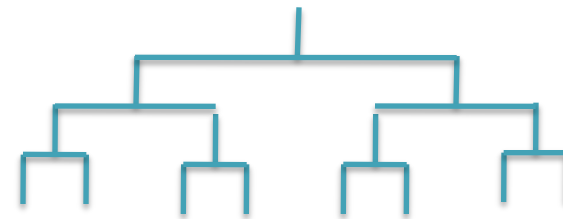
Alignment & Variations



Differential Analysis

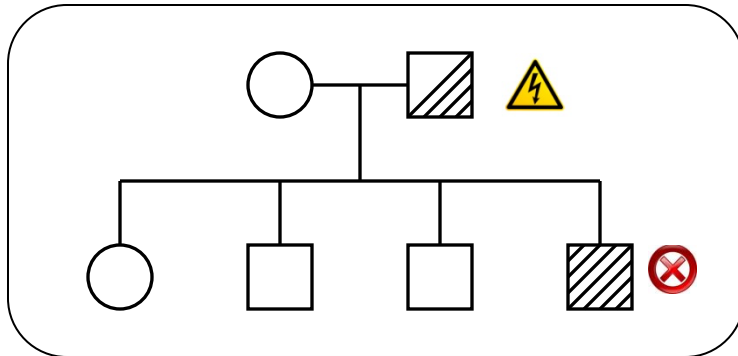


Phylogeny & Evolution



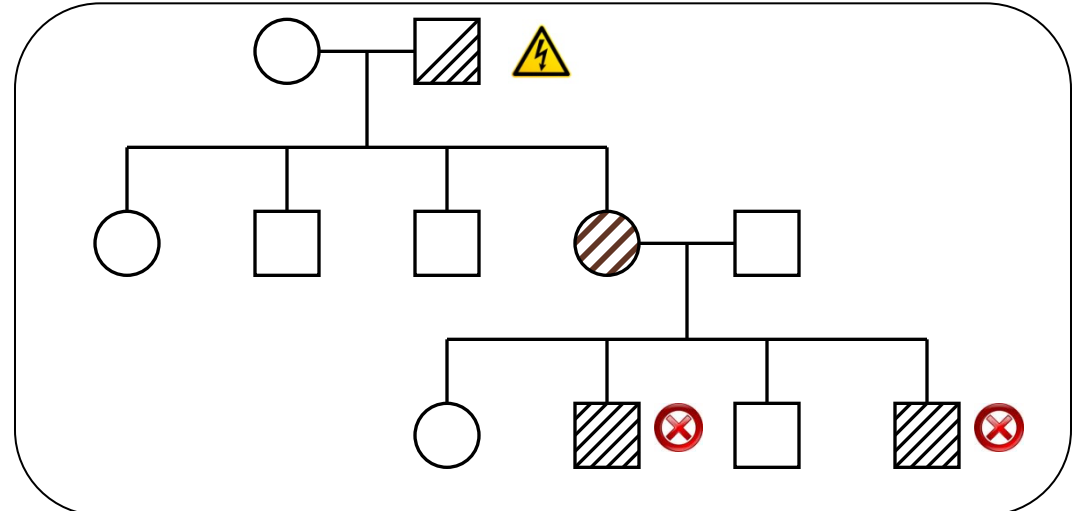
Unified Model of Autism

Sporadic Autism



De novo mutations of high penetrance contributes to autism, especially in families with lower risk than in families at higher risk.

Familial Autism



Legend



Sporadic mutation

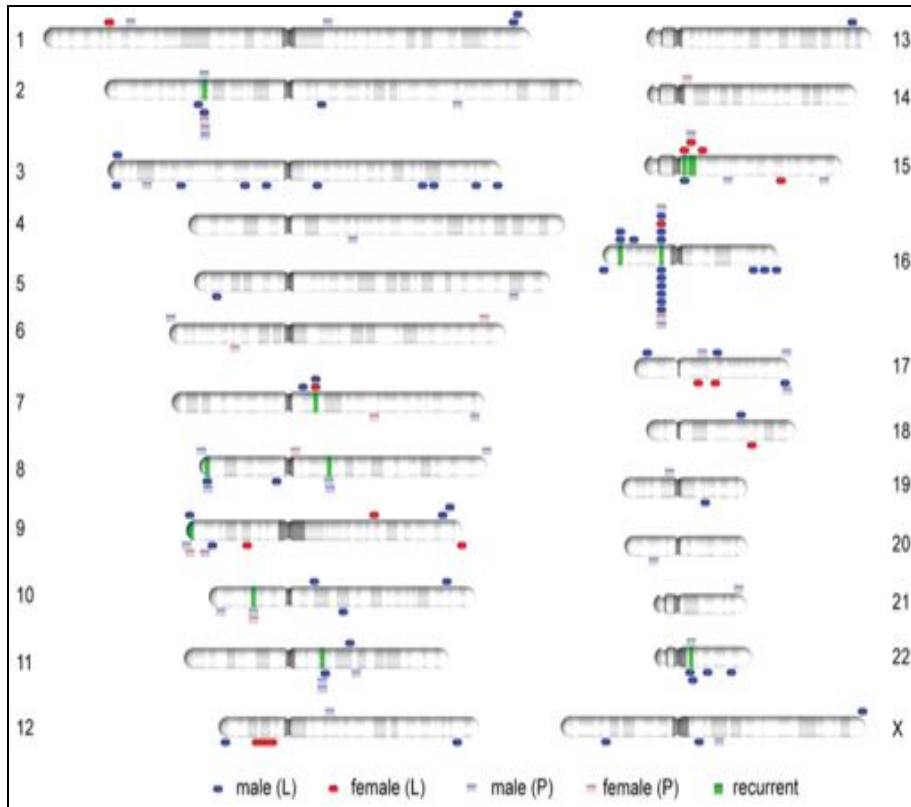


Fails to procreate

A unified genetic theory for sporadic and inherited autism

Zhao et al. (2007) *PNAS*. 104(31)12831-12836.

Autism and de novo CNVs



CNV analysis of Simons Simplex Collection

- CGH arrays of 510 family quads
- 94 total de novo CNVs discovered

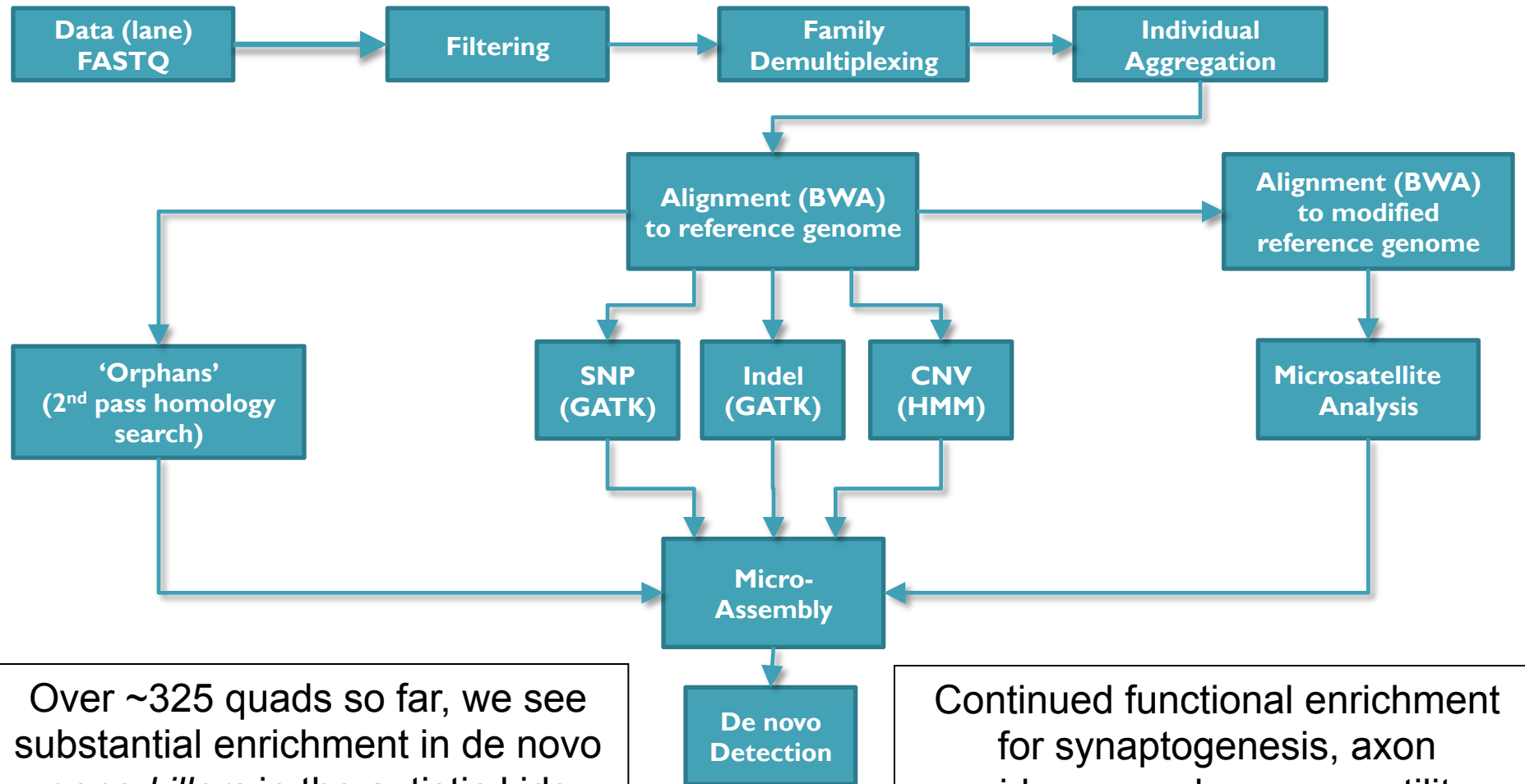
De novo CNVs enriched in autistic children

- 4:1 ratio in autistic kids relative to their non-autistic siblings
- Some recurrence at genes related to other psychiatric conditions

	Counts of De Novo Events			Children with De Novo Events			Frequency in Children		
	Combined	Del	Dup	Combined	Del	Dup	Combined	Del	Dup
aut	75	46	29	68	44	27	7.9%	5.1%	3.1%
sib	19	9	10	17	8	9	2.0%	0.9%	1.0%

Rare de novo and transmitted copy-number variation in autism spectrum disorders.
 Levy et al. (2011) *Neuron*. 70:886-897.

Exome Sequencing Pipeline



Over ~325 quads so far, we see substantial enrichment in de novo *gene-killers* in the autistic kids

Continued functional enrichment for synaptogenesis, axon guidance, and neuron motility.

Assessing the role of de novo gene-killers in the incidence of autism

Iossifov et al. (2012) In preparation

Scalpel: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz



- Use assembly techniques to identify complex variations from short reads
 - Improved power to find indels
 - Trace candidate haplotypes sequences as paths through assembly graphs



Ref: ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

Father: ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

Mother: ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

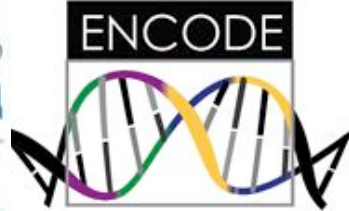
Sib: ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

Aut(1): ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

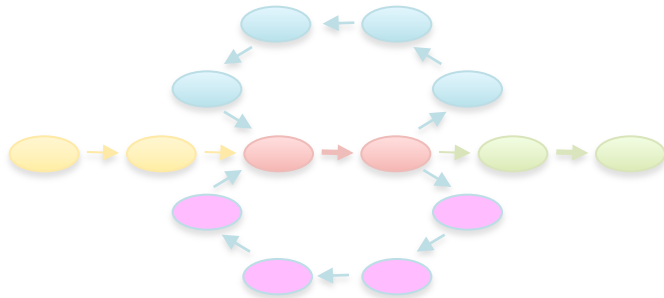
Aut(2): ...TCAGAACAGCTGGATGAGATCTTACC-----CCGGGAGATTGTCTTTGCCCGGA...

6bp heterozygous indel at chr13:25280526 ATP12A

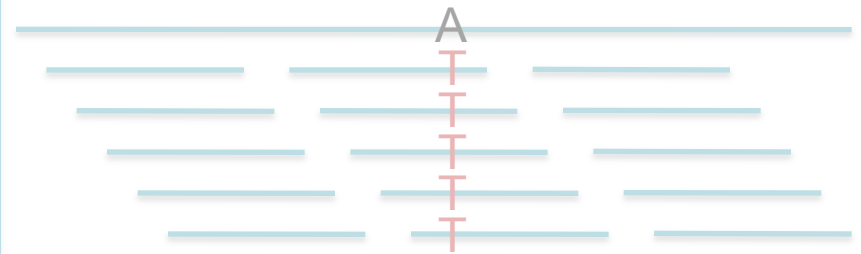
The rise of mega-genomics



De novo Assembly



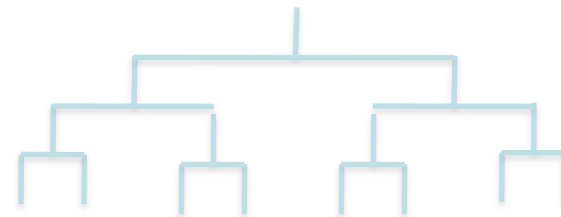
Alignment & Variations



Differential Analysis



Phylogeny & Evolution

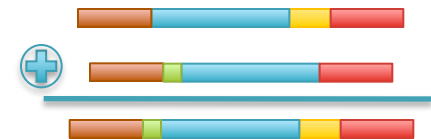
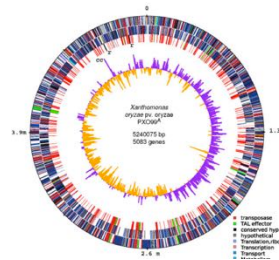


Summary

I'm focused on the intersection of the most significant biology, biotechnology, and compute technology

We are entering the era of mega-genomics

- Explosion in digital traits and measurements
- Parallel systems essential for analyzing large data sets
- Algorithms and machine learning to squeeze insight out of diverse data types
- Collaborations with biologists and visual informatics systems to help execute experiments & interpret results

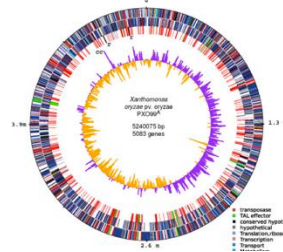


Acknowledgements



Giuseppe Narzisi
Mitch Bekritsky

Ivan Iossifov
Wigler Lab



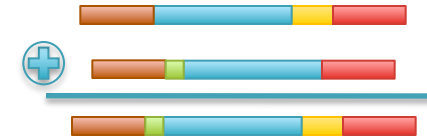
Hayan Lee
Matt Titmus
James Gurtowski

Ware Lab
McCombie Lab

Adam Phillippy (NBACC)
Sergey Koren (NBACC)



DOE Systems Biology Knowledgebase



Paul Baranay (CSHL/ND)

Scott Emrich (ND)
Steven Salzberg (JHU)
Mihai Pop (UMD)



Thank You!

<http://schatzlab.cshl.edu/apply>
[@mike_schatz](#)

