Scalable solutions for 2nd and 3rd gen sequencing

Michael Schatz

March 29, 2012
NYU HiTS Series

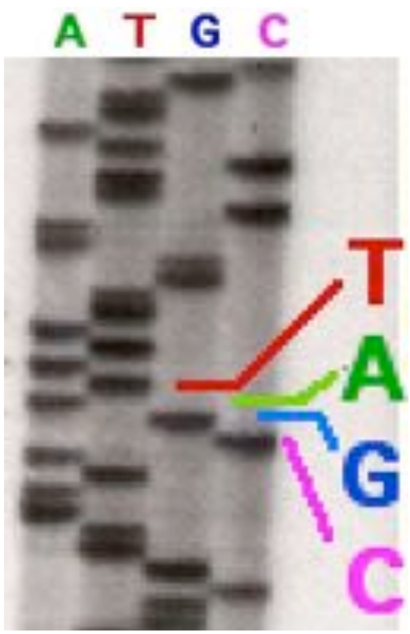
@mike_schatz

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

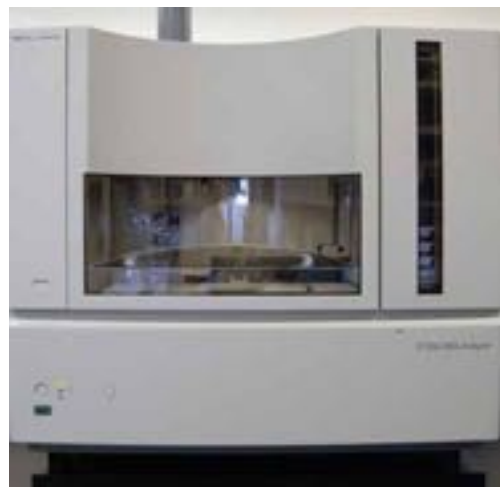
Advances in Sequencing: Zeroth, First, Second Generation



1970s: 0th Gen

Radioactive Chain Termination

5000bp / week



1980s-1990s: 1st Gen

Automated Capillary Sequencing

384kbp / day

2000s: 2nd Gen

Pyrosequencing, SOLiD Sequencing-by-Synthesis

1Gbp+ / day

Advances in Sequencing: Now Generation Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

>60Gbp / day

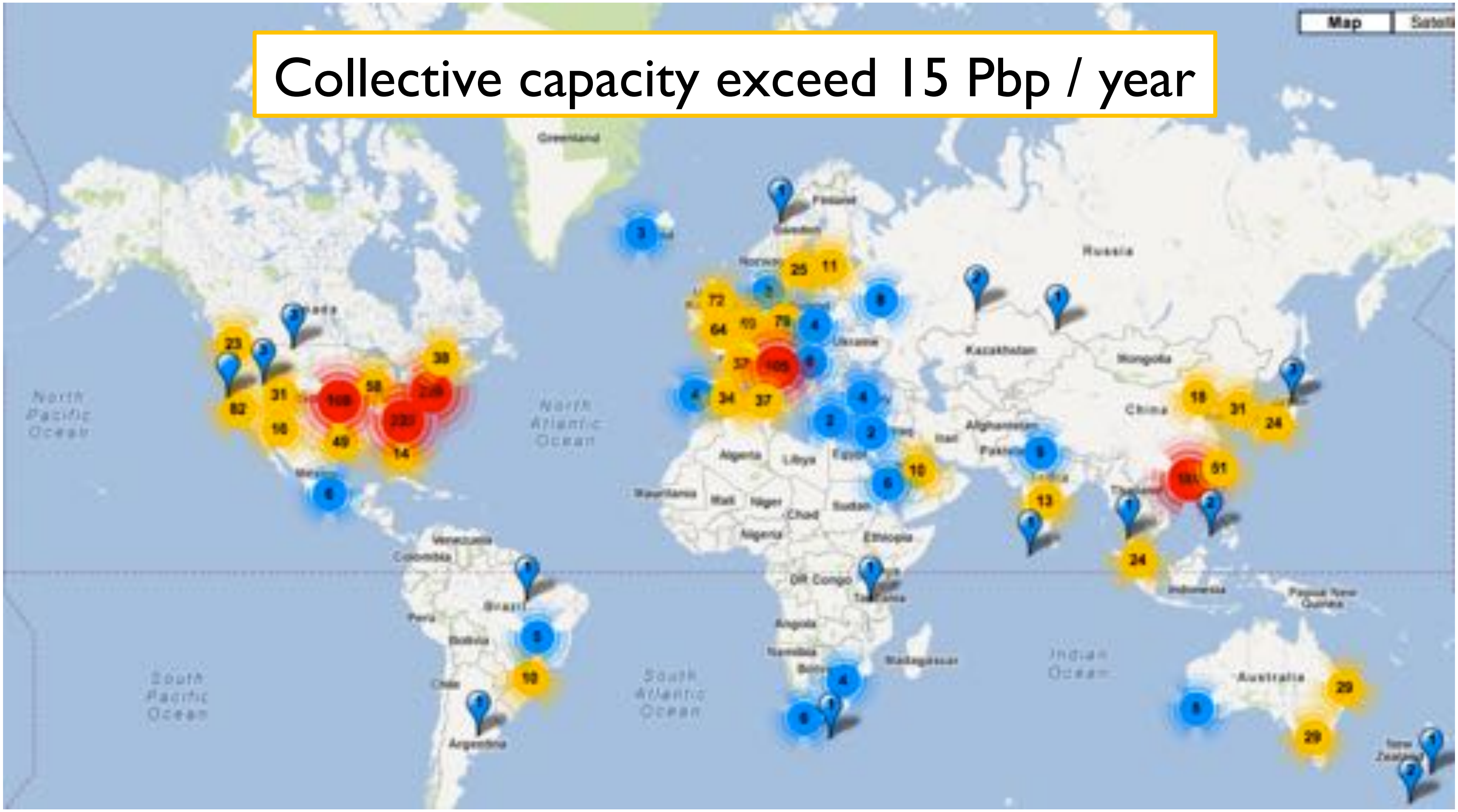
Ion Proton
Postlight Sequencing

>100Gbp / day

Oxford Nanopore
Nanopore sensing

Many GB / day

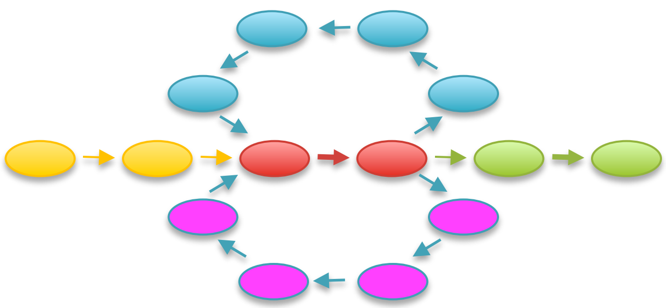
Sequencing Centers



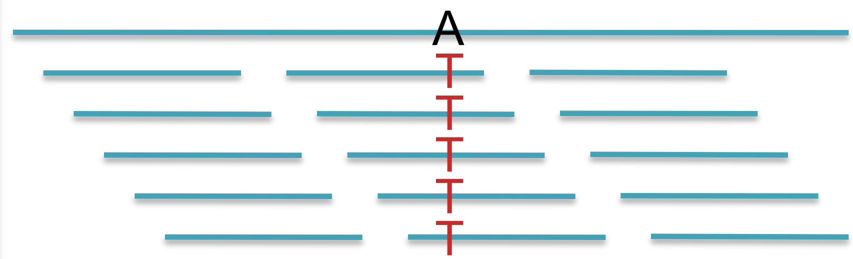
Next Generation Genomics: World Map of High-throughput Sequencers
http://pathogenomics.bham.ac.uk/hts/

The rise of mega-genomics

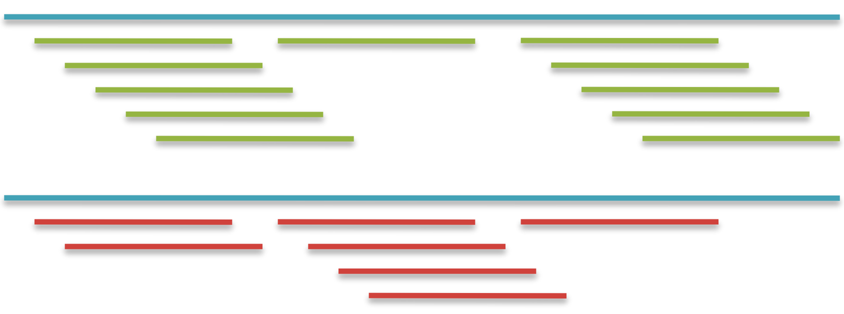
De novo Assembly



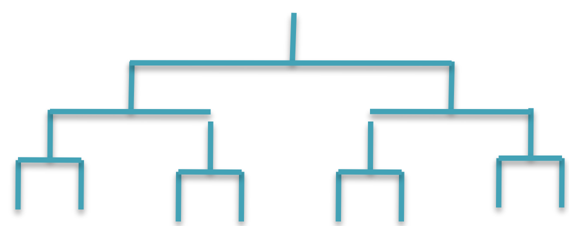
Alignment & Variations



Differential Analysis

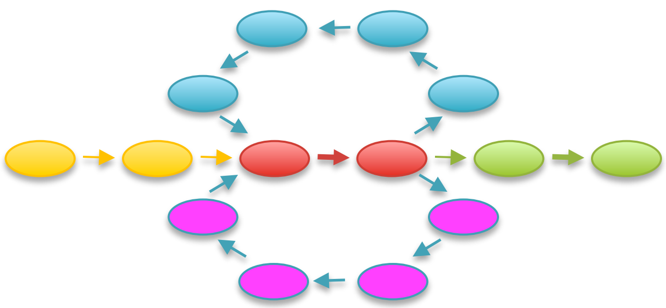


Phylogeny, Evolution, and Modeling

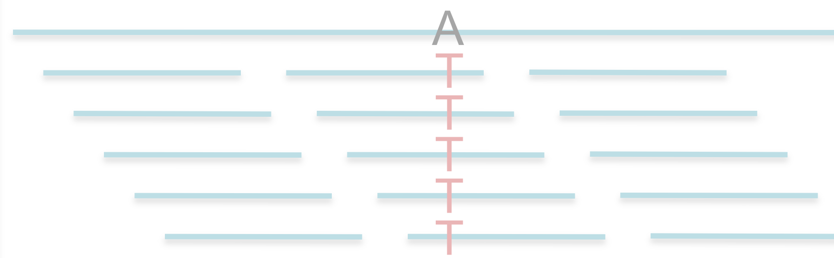


The rise of mega-genomics

De novo Assembly



Alignment & Variations

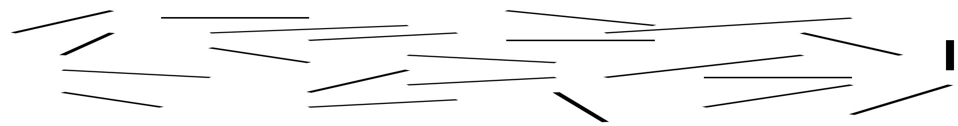


Differential Analysis

Phylogeny, Evolution, and Modeling

Assembling a Genome

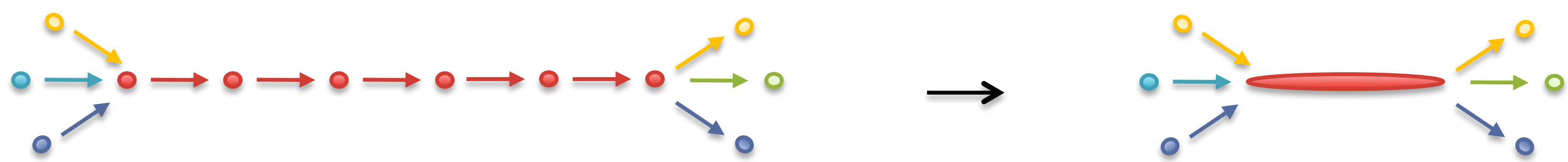
1. Shear & Sequence DNA



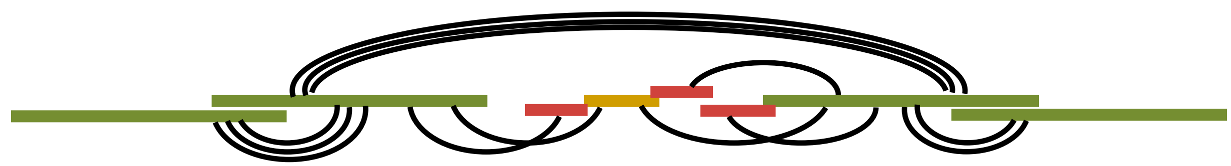
2. Construct assembly graph from overlapping reads

...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...

3. Simplify assembly graph



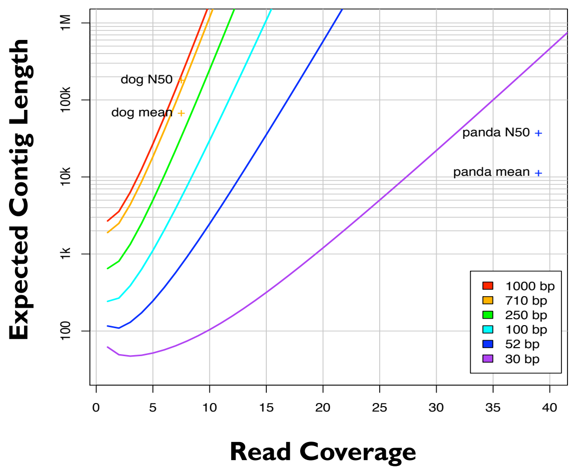
4. Detangle graph with long reads, mates, and other links



De novo genome assembly: what every biologist should know
Monya Baker (2012) *Nature Methods*. 9:333-337.

Ingredients for a good assembly

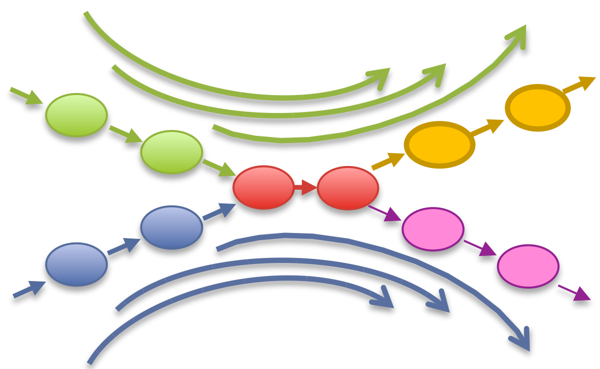
Coverage



High coverage is required

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

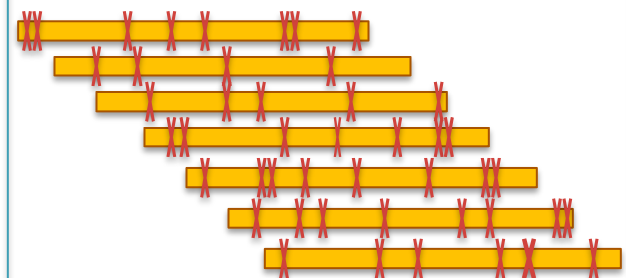
Read Length



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

Current challenges in de novo plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology.* In Press.

Hybrid Sequencing

Illumina
Sequencing by Synthesis

High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

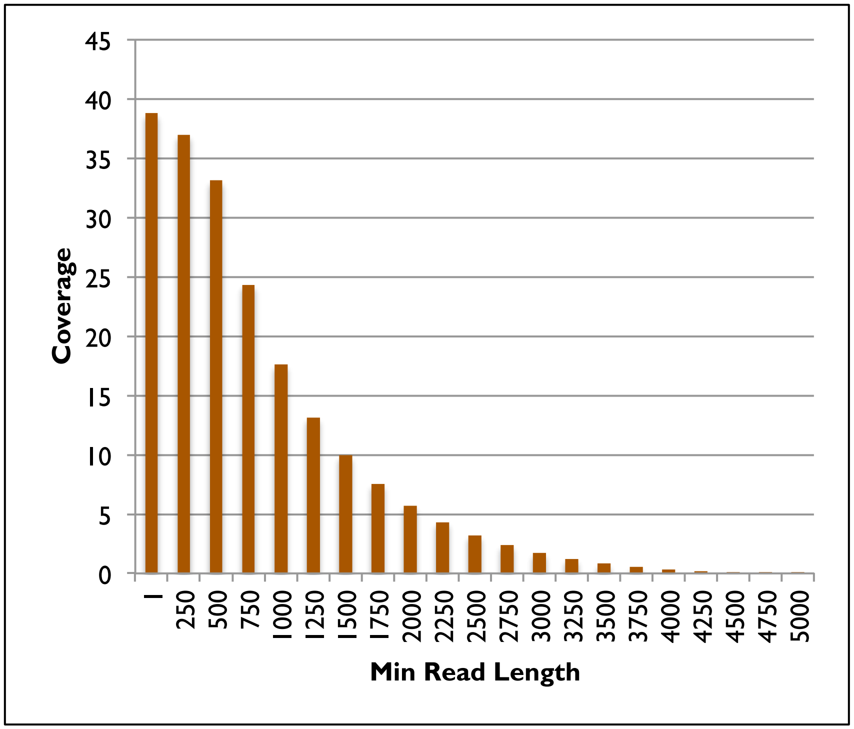
Pacific Biosciences
SMRT Sequencing

Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (10kbp+)

SMRT Sequencing Data

Yeast
(12 Mbp genome)

65 SMRT cells
734,151 reads after filtering
Mean: 642.3 +/- 587.3
Median: 553 Max: 8,495




```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| |||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

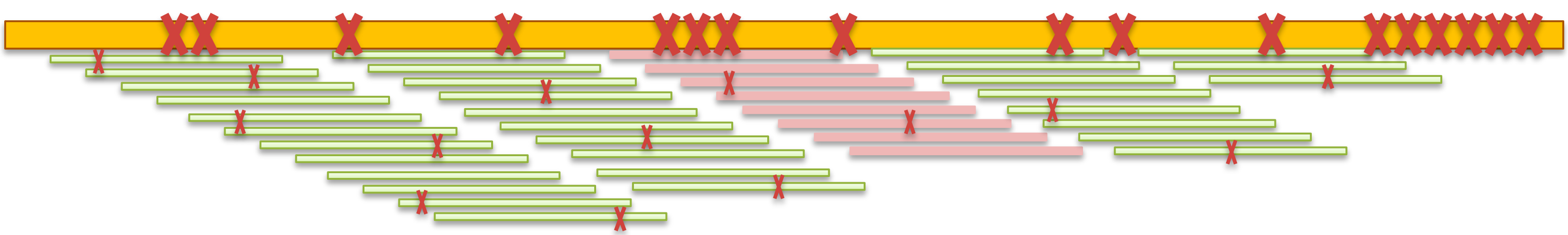
Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy: 83.7%, 11.5% insertions, 3.4% deletions, 1.4% mismatch

PacBio Error Correction

http://wgs-assembler.sf.net

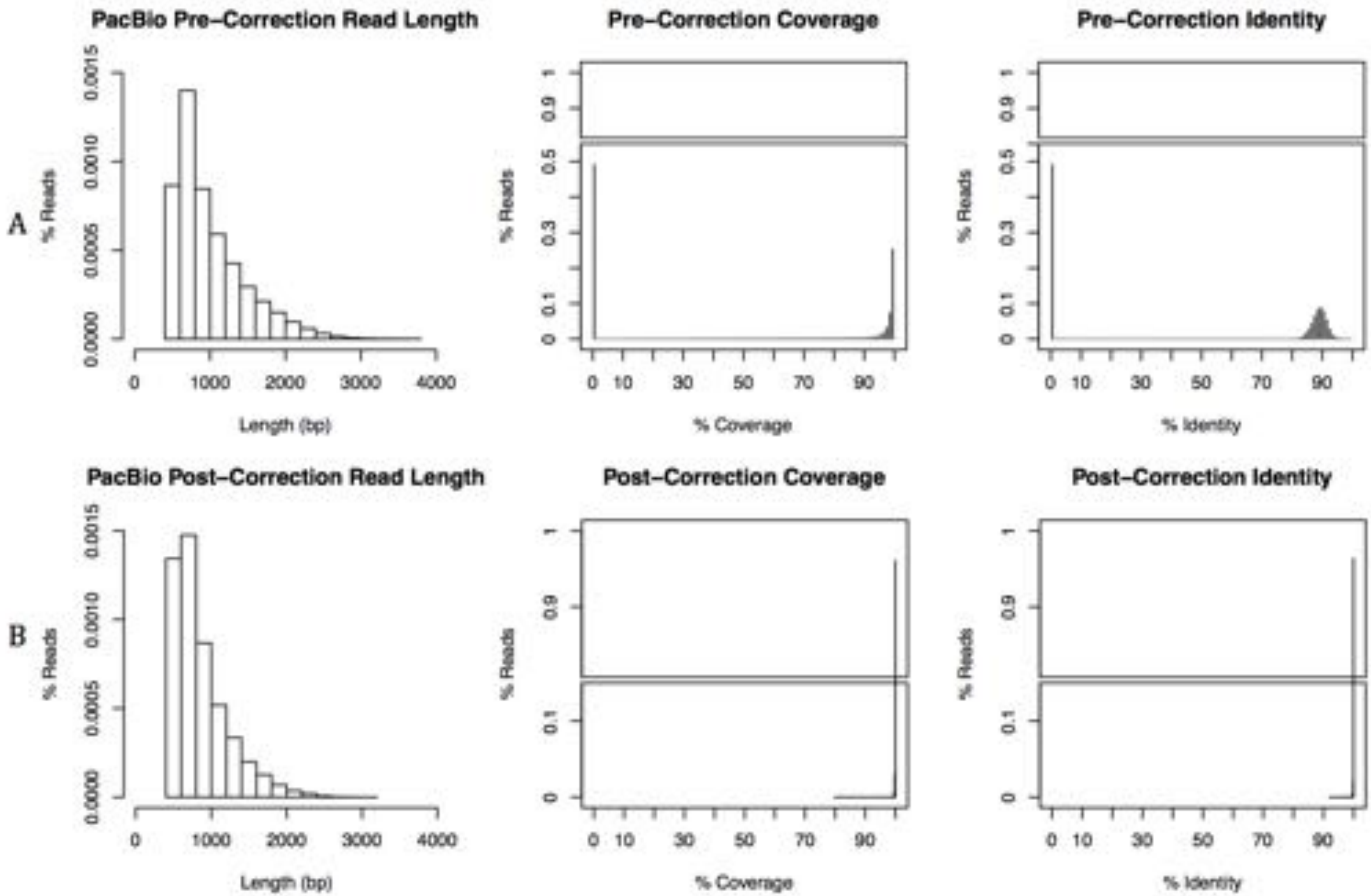
1. Correction Pipeline
   1. Map short reads (SR) to long reads (LR)
   2. Trim LRs at coverage gaps
   3. Compute consensus for each LR

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, WR, Jarvis, ED, Phillippy, AM. (2012) *Under Review*

Error Correction Results

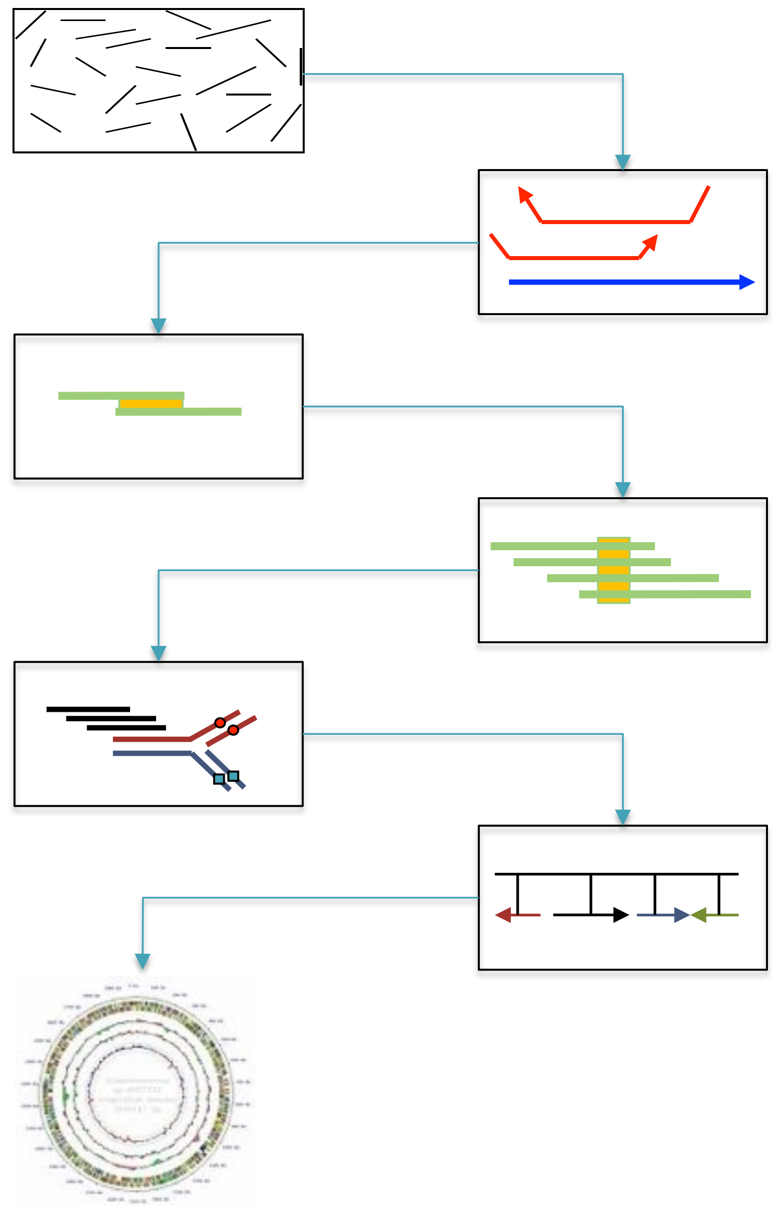


Correction results of 20x PacBio coverage of E. coli K12 corrected using 50x Illumina

Celera Assembler

http://wgs-assembler.sf.net

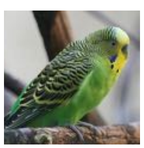
1. Pre-overlap
   – Consistency checks
2. Trimming
   – Quality trimming & partial overlaps
3. Compute Overlaps
   – Find high quality overlaps
4. Error Correction
   – Evaluate difference in context of overlapping reads
5. Unitigging
   – Merge consistent reads
6. Scaffolding
   – Bundle mates, Order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



SMRT-Assembly Results







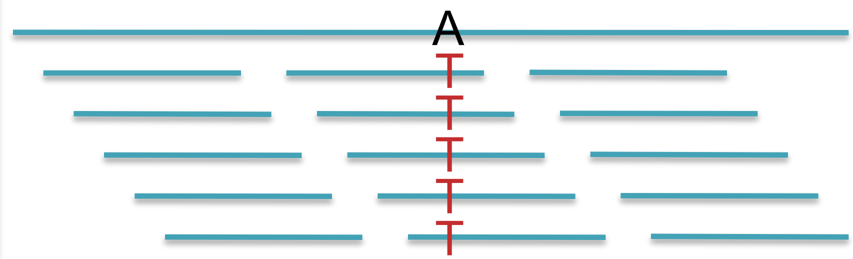
Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.76%) *
(median: 747 max: 3 068)	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%) *
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
(median: 1 217 max: 14 901)	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[a]		5 452 251	23	653 382	402 041
S. cerevisiae S228c	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
(median: 674 max: 5 994)	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%) *
Melopsittacus undulatus	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
(median 997, max 13 079)	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
*** **Also useful for transcriptome, repeat, and other analysis** ***

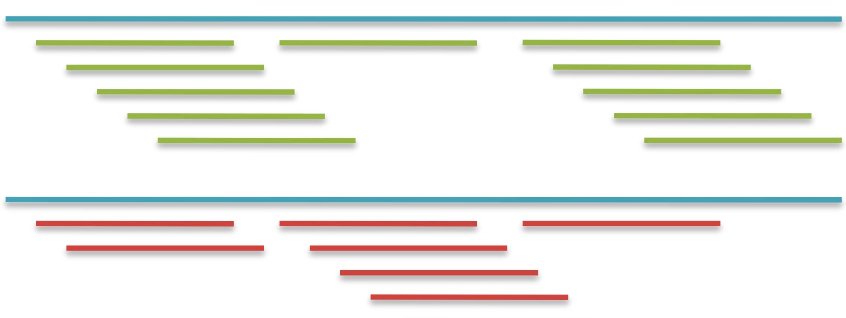
The rise of mega-genomics

De novo Assembly

Alignment & Variations

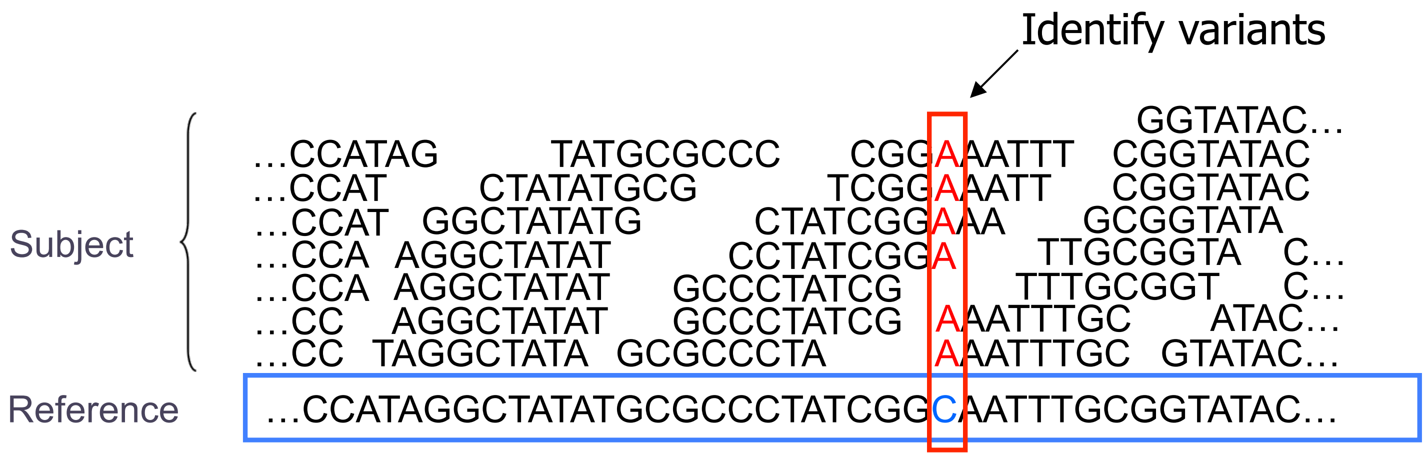


Differential Analysis



Phylogeny, Evolution, and Modeling

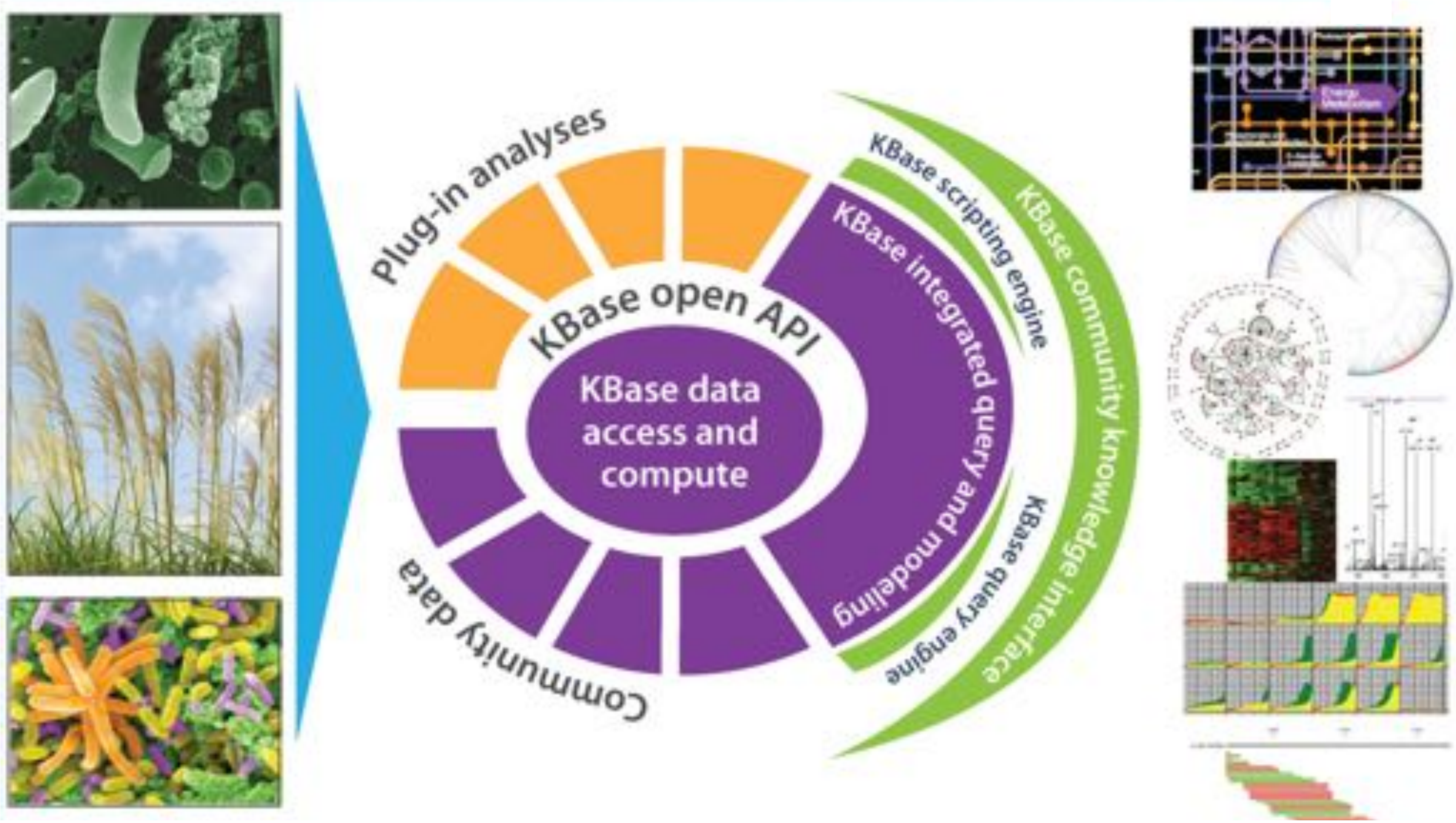
Short Read Mapping



- Given a reference and many subject reads, report one or more "good" end-to-end alignments per alignable read
  - Fundamental computation for many assays
    - RNA-seq Methyl-seq FAIRE-seq
    - ChIP-seq Dnase-seq Hi-C-seq

- Desperate need for scalable solutions
  - Single human requires >1,000 CPU hours / genome
  - **1000 hours * 1000 genomes = 1M CPU hours / project**

The DOE Systems Biology Knowledgebase



http://kbase.us: Predictive Biology in Microbes, Plants, and Meta-communities

Hadoop MapReduce

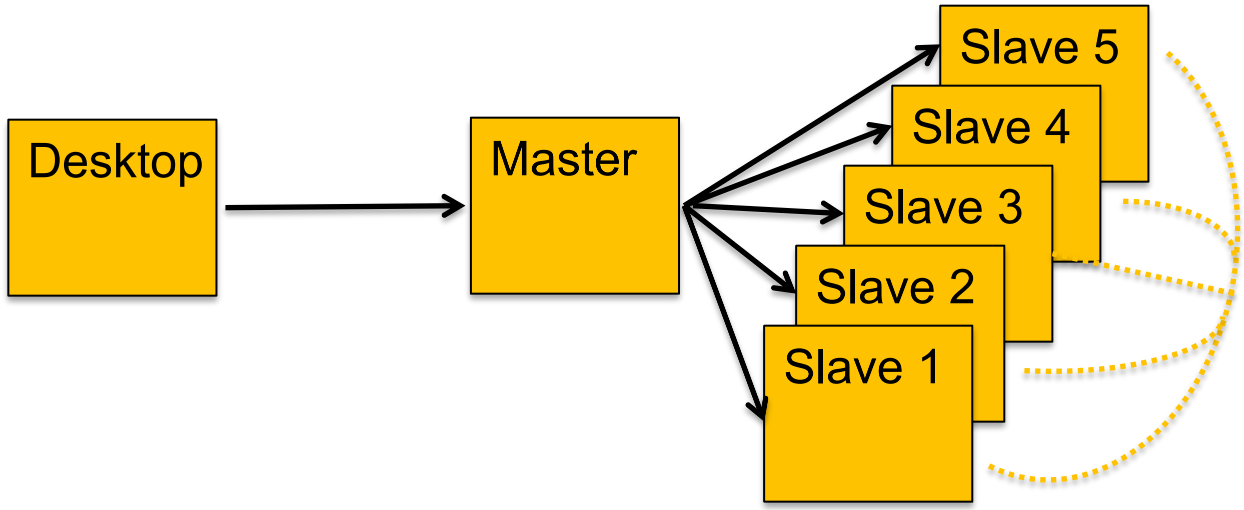
http://hadoop.apache.org

- 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

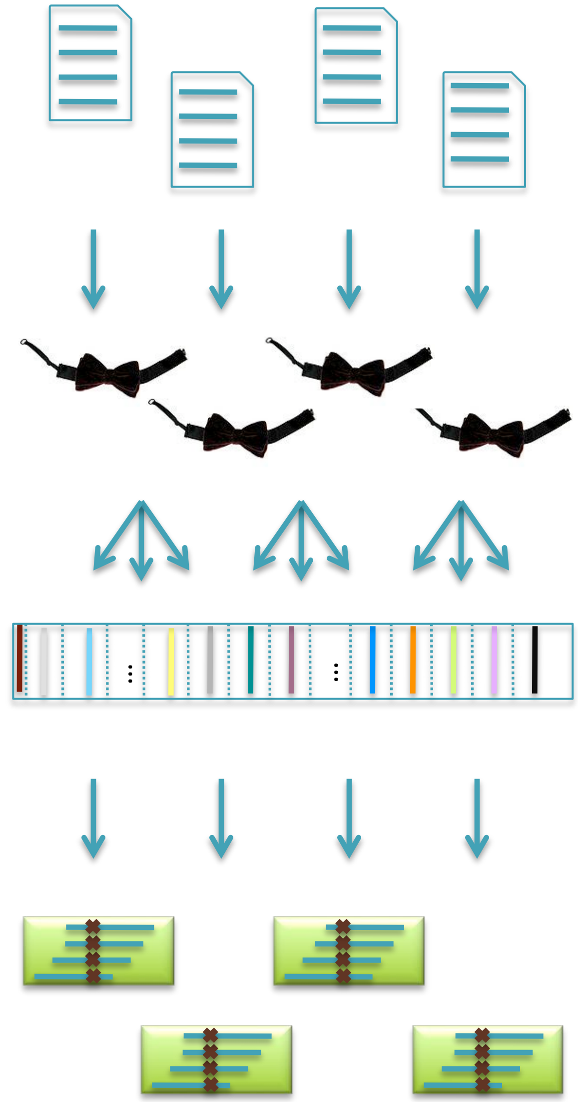
System Architecture



- 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

Hadoop for NGS Analysis

CloudBurst

Highly Sensitive Short Read Mapping with MapReduce

100x speedup mapping on 96 cores @ Amazon

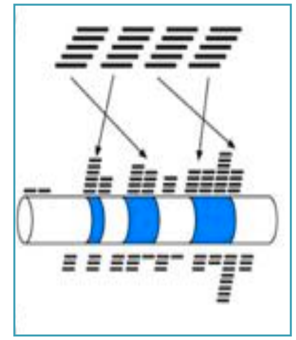
http://cloudburst-bio.sf.net

(Schatz, 2009)

Myrna

Cloud-scale differential gene expression for RNA-seq

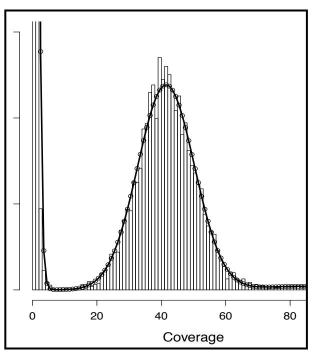
Expression of 1.1 billion RNA-Seq reads in ~2 hours for ~$66



(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

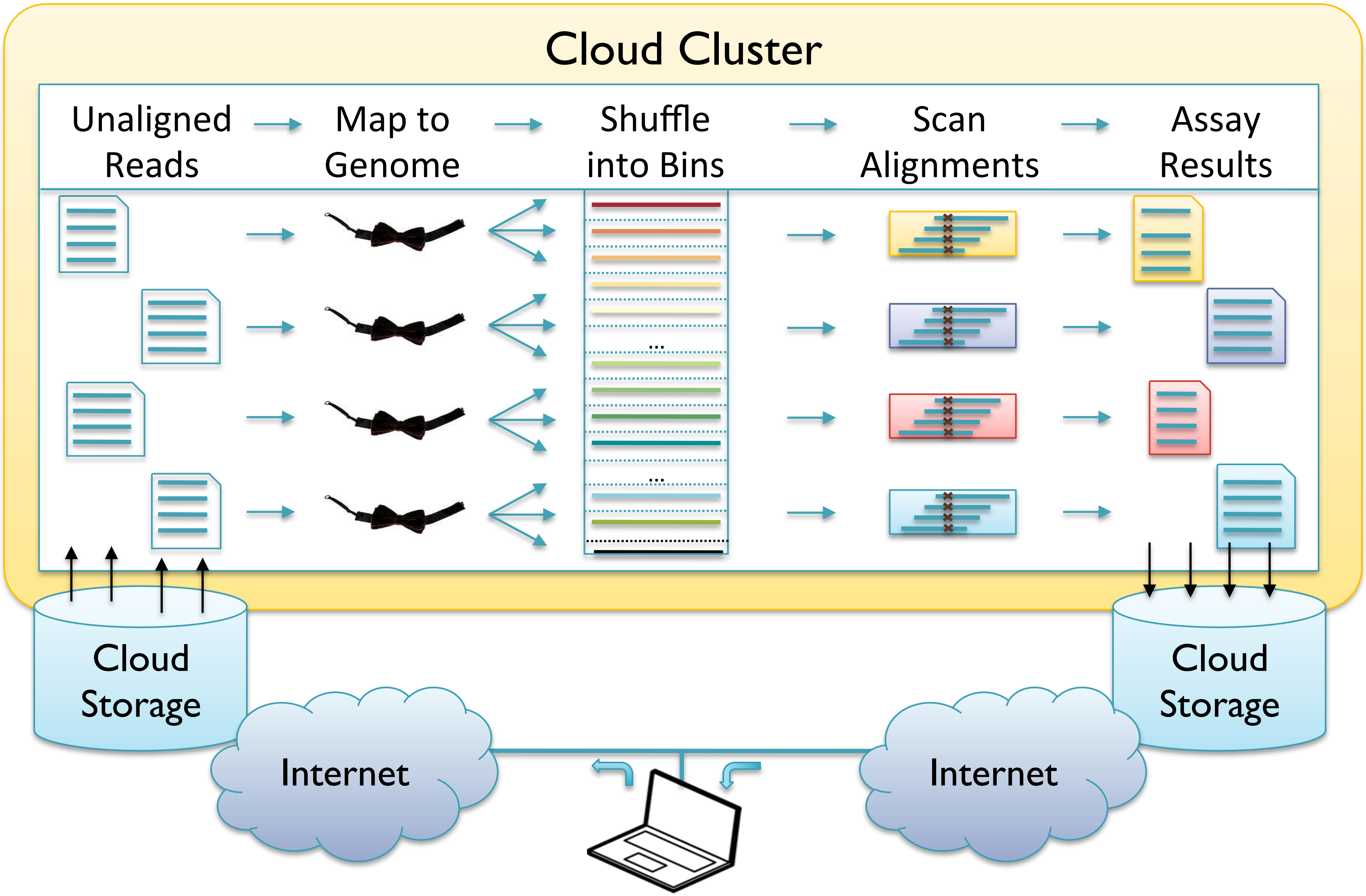
Construct the BWT of the human genome in 9 minutes

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

(Menon, Bhat, Schatz, 2011)

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

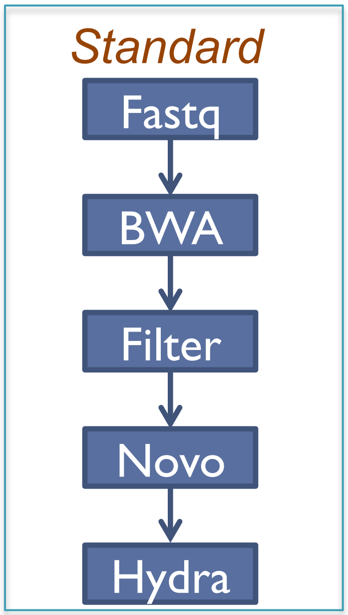
Map-Shuffle-Scan for Genomics

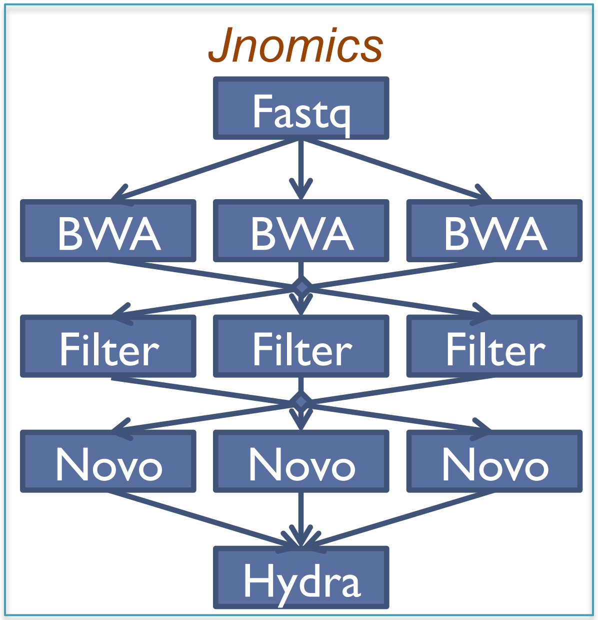


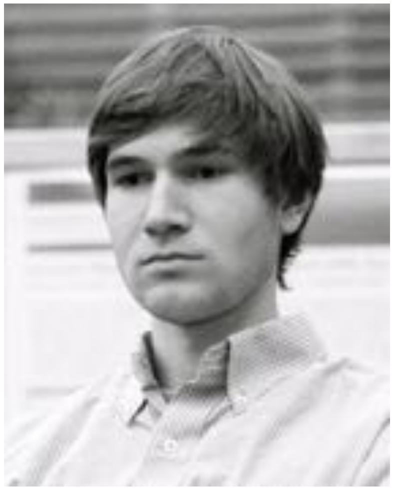
Cloud Computing and the DNA Data Race.
Schatz, MC, Langmead B, Salzberg SL (2010) *Nature Biotechnology.* **28**:691-693

Jnomics: Cloud-scale genomics

Matt Titmus, 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

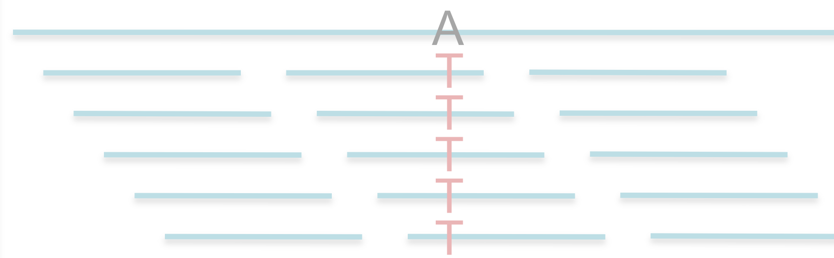
Answering the demands of digital genomics
Titmus, M.A.., Schatz, M.C.. (2012) *Under Review*

The rise of mega-genomics

GENOME 10K ENCODE HMP

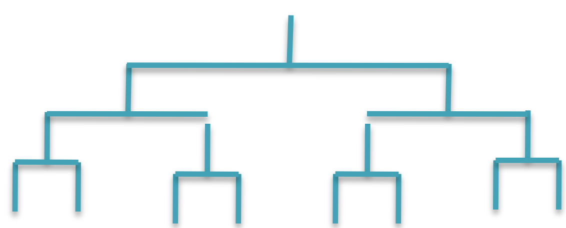
De novo Assembly

Alignment & Variations



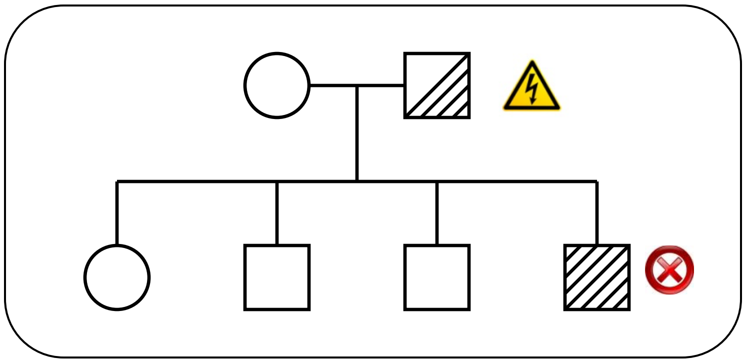
Differential Analysis

Phylogeny, Evolution, and Modeling



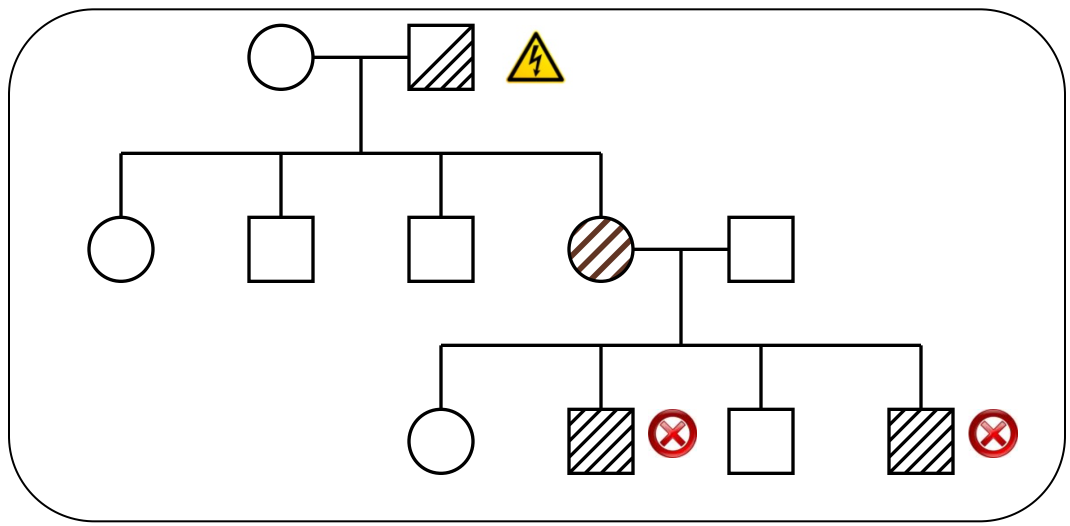
Unified Model of Autism

Sporadic Autism

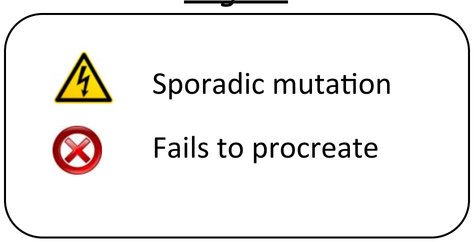


De novo mutations of high penetrance contributes to autism, especially in low risk families with no history of autism.

Familial Autism

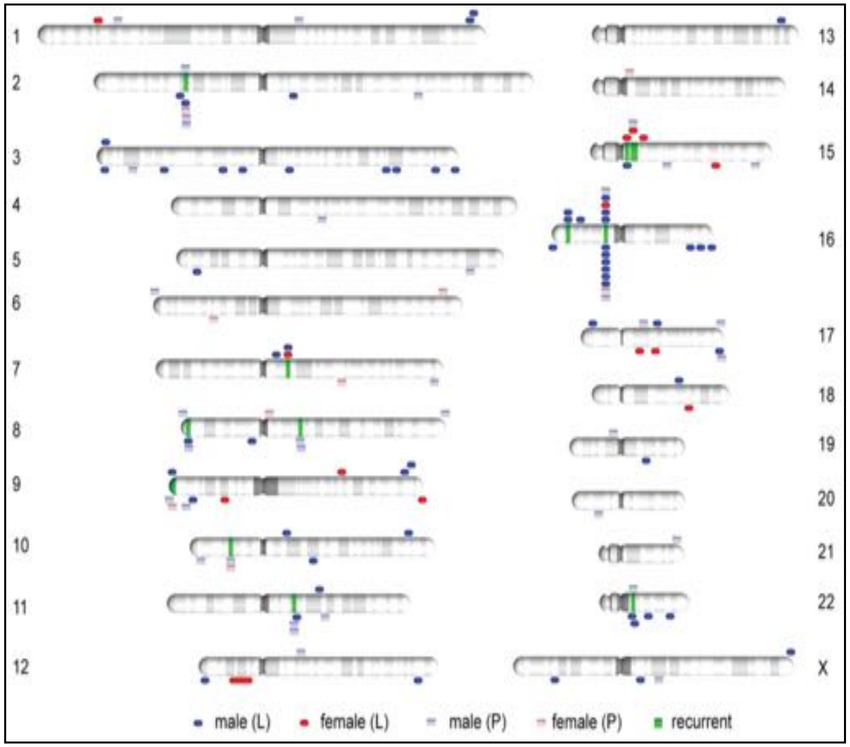


Legend



A unified genetic theory for sporadic and inherited autism
Zhao *et al.* (2007) *PNAS. 104(31)12831-12836.*

Autism and de novo CNVs



Analysis of Simons Simplex Collection

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

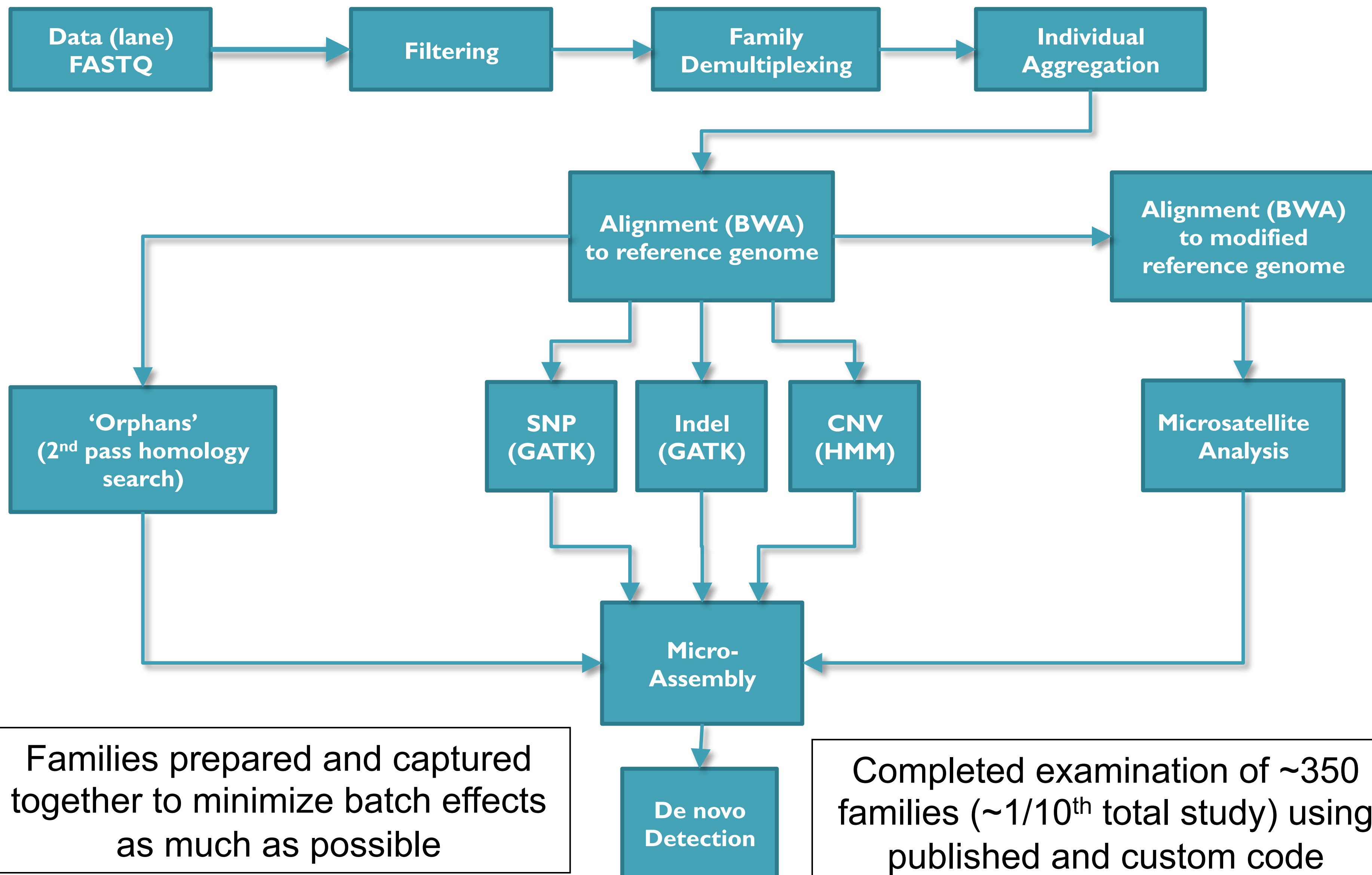
De novo CNVs are more common in autistic children

- 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



Families prepared and captured together to minimize batch effects as much as possible

Completed examination of ~350 families (~1/10th total study) using published and custom code

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

De novo Genetics of Autism

- In 343 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1 (432:396)
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations

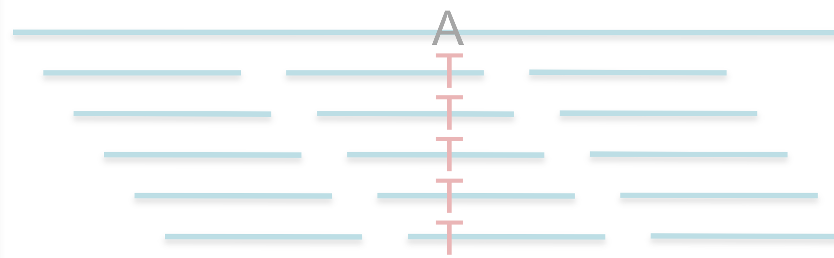
- Observe strong overlap with the 842 genes known to be associated with fragile X mental retardation.
  - These genes relate to neuron and brain development
  - Suggest these genes are under strong purifying selection and we hypothesize particularly dosage sensitive

Exome sequence analysis of simplex families with children on the autism spectrum
Iossifov *et al. (2012)* Under review

The rise of mega-genomics

De novo Assembly

Alignment & Variations



Differential Analysis

Phylogeny, Evolution, and Modeling

Mega-Genomics Challenges

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

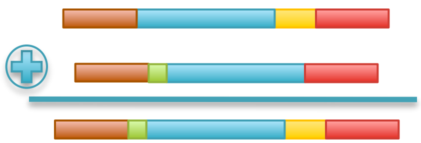
- Technology will continue to push the frontier
- Measurements will be made *digitally* over large populations, at extremely high resolution, and for diverse applications

Rise in Quantitative and Computational Demands

1. *Experimental design*: selection, collection & metadata
2. *Observation*: measurement, storage, transfer, computation
3. *Integration*: multiple samples, assays, analyses
4. *Discovery*: visualizing, interpreting, modeling

Ultimately limited by the human capacity to execute extremely complex experiments and interpret results

Acknowledgements

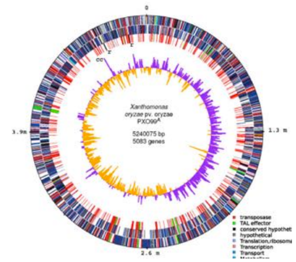


McCombie Lab

Adam Phillippy (NBACC)
Sergey Koren (NBACC)

Paul Baranay (CSHL/ND)
Scott Emrich (ND)
Steven Salzberg (JHU)
Mihai Pop (UMD)

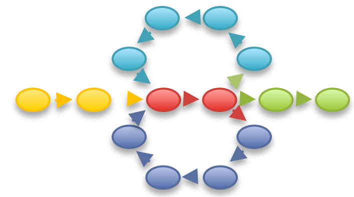
National Human Genome Research Institute



James Gurtowski
Matthew Titmus

Ware Lab
KBase Members

KBASE predictive biology
DOE Systems Biology Knowledgebase



Giuseppe Narzisi
Mitch Bekritsky

Ivan Iossifov
Wigler Lab

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

Thank You!

http://schatzlab.cshl.edu/apply

@mike_schatz