Genome Center Rebels:

Next-Gen Analysis Outside the Machine

COFACTOR
GENOMICS

Cofactor Genomics

Feedback-driven optimization of data generation

Experimental design

Molecular biology

DNA sequencing

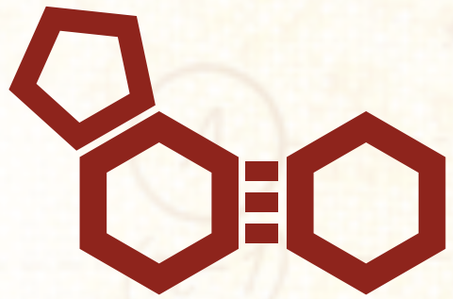
Data analysis

Visualization

Advanced visualization tools enable complex design

COFACTOR GENOMICS

The Transition



- Unlimited Sequencing/Reagent Resources
- 1000+ CPU cluster, 100s of TBs of Disk
- Platform LSF job scheduler
- IT staff
- Flexible deadlines
- 99% Human Re-sequencing

- Tight budget with no wiggle room
- Modest compute and storage resources
- Manual job execution
- ...No way.
- Customers with grant deadlines!
- 1% human resequencing

COFACTOR GENOMICS

Top Requested Applications

Fragment Whole genome characterization by single-pass shotgun sequencing of fragments from total DNA, PCR products, etc

Paired-end Whole genome characterization by shotgun sequencing from both ends of DNA fragments with ~200bp inserts

RNA-Seq Quantitative transcriptome profiling by sequencing cDNAs constructed from messenger RNA isolated from total RNA

miRNA Discovery & quantitation of novel microRNAs and isoforms by sequencing cDNAs of microRNAs isolated from total RNA

Bisulfite Genome methylation profiling by sequencing DNA fragments bisulfite treated to convert non-methylated C's into U's

ChIP-Seq Discovery & quantitation of protein-DNA interactions by sequencing DNA from immunoprecipitations

COFACTOR GENOMICS

First 10 Months of Libraries

50	RNA-seq
39	Genomic
24	small RNA
20	Genomic Reduced Representation Sequencing (RRS)
19	ChIP-seqs of Transcripton Factors and Histone Mods
10	Pooled Patient/Crop PCR
6	Pooled Loci PCR
3	ChIP-seq of RNA-binding proteins
3	Bisulphite converted **Not currently recommended!*
3	DNAse 1 Hypersensitivity (DHS)
2	"PCR-free" genomic for 2 high AT genomes
1	Pooled Bacterial Genomes

---

180 **Total Libraries,** most Paired-End if relevant

COFACTOR GENOMICS

Recurring Analysis Issues & Idioms

Too Many Tools A plethora of substitutable tools, few of which are worth using, such as: MAQ, Mosaik, SOAP, SHRiMP, BowTie, NovoAlign or Velvet, Euler-SR, Edena, All-Paths, AMOScmp-shortReads, AbySS

Poor Algorithms Single-threaded, compromise-accuracy-for-performance, memory-hog applications like MAQ, Velvet, Euler-SR, ELAND...

Poor Data Formats Giant uncompressed TXT files from Illumina & AB, useless design-by-committee formats like SRF, non-standardized formats like GFF

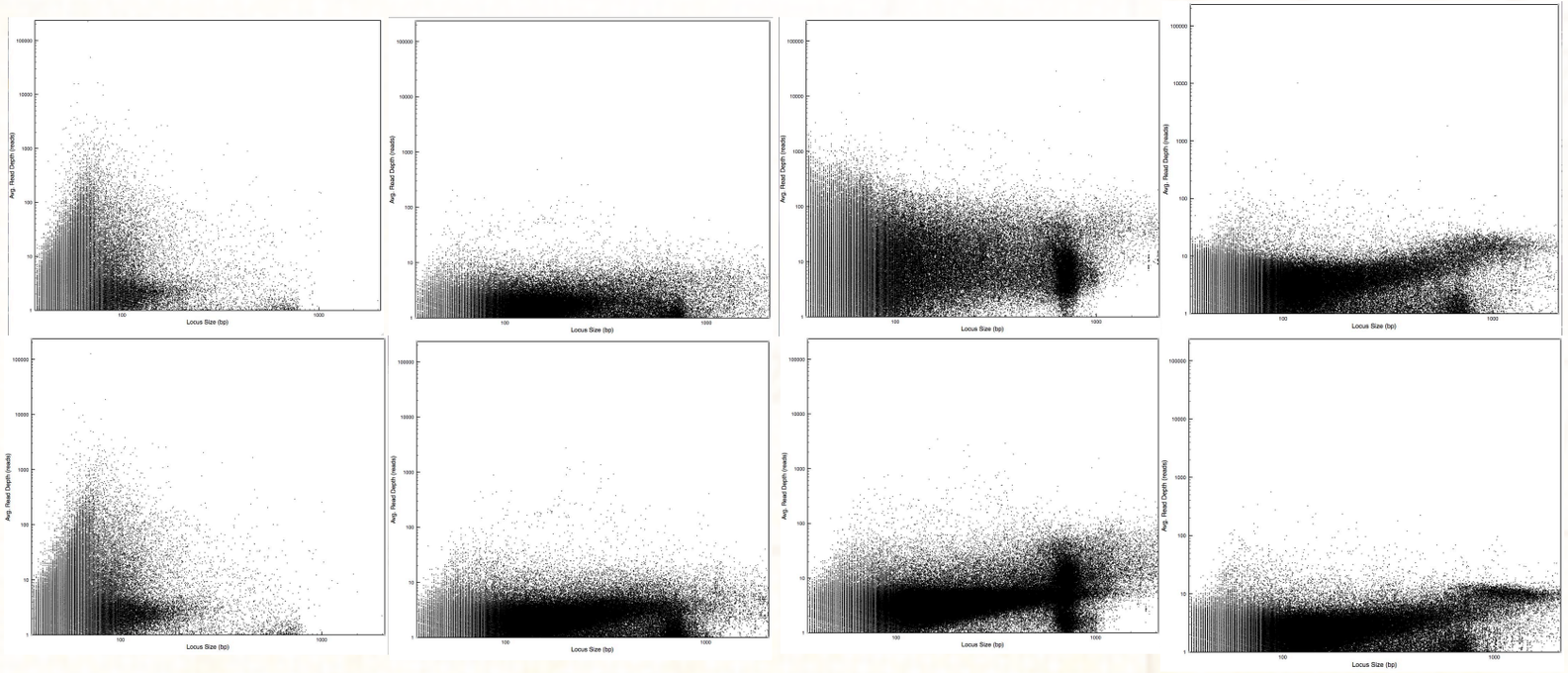
Recurring Intermediates Base-by-base coverage, lists of intervals (like annotations or clusters), base-by-base nucleotide count and quality, FASTA, alignments

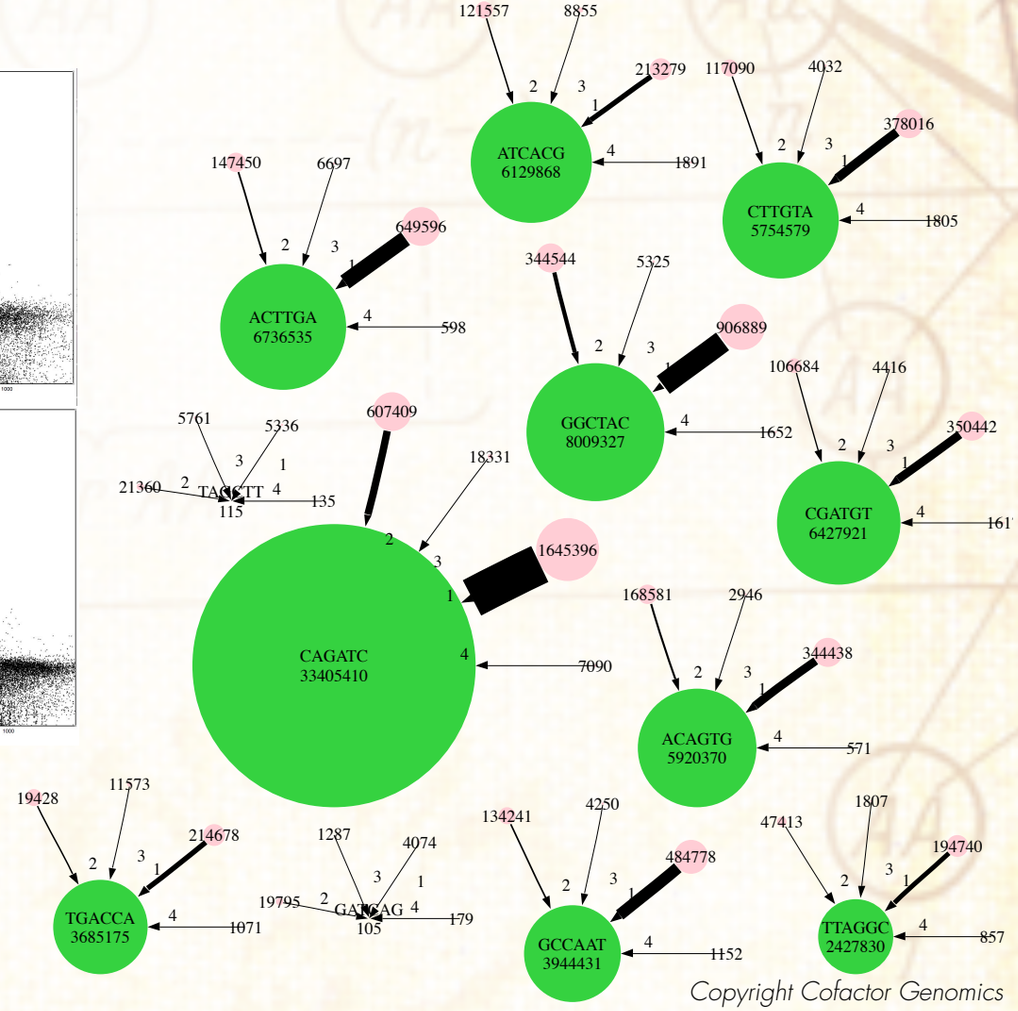
Recurring Idioms Base-by-base whole genome iteration, annotation directed base-by-base iteration, assemblies of non-mapping reads and mapping of original reads to assemblies...

COFACTOR GENOMICS

Analysis on a Shoestring - Illumina Barcoded ChIP

- Re-write Illumina barcode binning
  (saves 20MB -- an old 454 run's worth of data -- per lane)
- Compute base-by-base coverage from alignments
- Iterate through bases to find regions of contiguous coverage





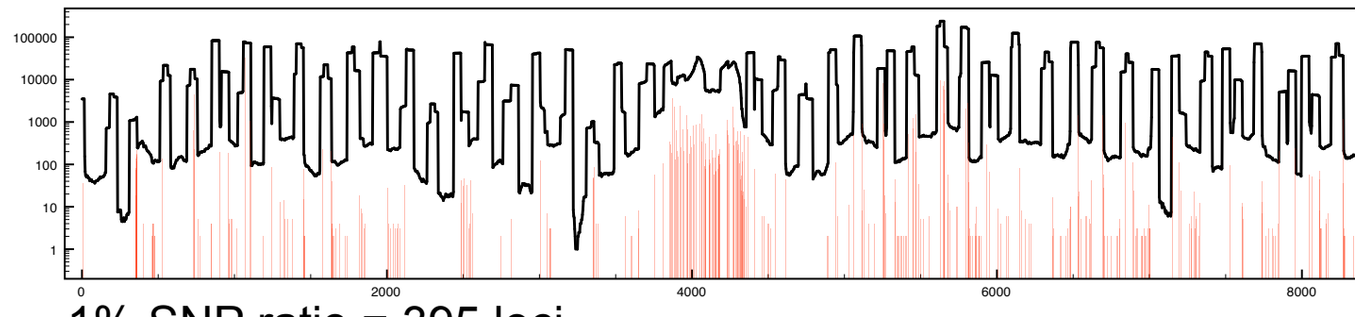
DTTT 1 6 22 21
CCCH 19 9 11 1
HTTT 1 12 10 14
CCCV 2 40 12 1
AC[illegible]
AAAA 3 6 2 13
AAMW 9 12 2 1
BCCC 1 15 11 9
AAAA 3 3 4 9
GGGT 2 6 3 3
AAAA 5 12 4 6
CCCCC 40 3 9 17 4
TTTTW 16 15 26 12 2
BGGGG 1 40 5 17 3
TTTTT 14 14 16 8 4
AAAAA 17 4 6 9 4
ACCCGY 6 32 8 4 2 1
GGGKK 40 2 12 2 2
CCCHMM 16 25 4 1 2 1
AGGGG 7 15 40 5 34
GGGGG 26 40 6 3 8
CCCCM 4 40 4 6 2
GGGGGV 14 40 40 5 6 1
CGTTTT 16 3 40 26 14 5
TTTTTTT 28 40 40 6 40 24 6
AAGTTTT 8 17 6 27 14 40 40
AAGGGGG 9 7 5 40 40 40 14
DDTTTTTTTT 1 1 4 40 28 19 3 40 18 2
GGGGGGGGGG 3 40 37 40 2 4 40 13 40 5 4
GGGGGGGGGGGG 8 36 39 16 40 2 11 40 40 15 36 9 4
AGGGGGGGGGGG 4 6 36 32 40 40 7 40 40 40 39 4 4
CCCCCCCCCCHMY 15 27 29 29 40 19 40 11 28 4 1 2 2
AAAAAAAAAAAA 14 40 28 10 40 15 6 40 40 7 40 14 13
GGGGGGGGGGRR 29 14 28 32 40 3 4 40 39 26 40 2 2
AAAAAAAAAAAA 12 12 40 17 40 16 9 28 40 40 23 16 12
TTTTTTTTTTTT 21 4 32 32 40 4 4 40 40 15 38 7 11
AAAGGGGGGGGGGR 15 40 4 11 40 40 22 40 40 40 40 4 32 14 4 1
AAAAAAAAAAAAAAAAAAA 6 34 40 12 11 40 40 2 25 40 40 40 40 40 40 40 16 40 32 12 12
BCCCCCCCCCCCCCCCCCCCCCCCCCC 1 22 18 40 40 9 19 40 40 7 3 40 40 4 40 40 40 40 40 40 40 40 40 40 25 18 7 10
AAAAGGGGGGGGGGGGGGGGGGGGGGGGS 25 40 40 14 8 22 38 23 38 40 40 10 40 40 5 40 40 8 40 40 40 40 40 40 37 40 18 11 6 1
AAGGGGGGGGGGGGGGGGGGGGGGGGGG 12 4 22 40 40 4 34 40 40 40 40 40 40 7 3 40 40 40 40 40 40 40 40 40 28 31 40 32 10 4
AAAAAAAAAAAAAAAAAAAAAAAAAAAA 12 17 30 40 13 15 40 30 40 28 40 40 40 20 6 20 40 40 40 40 40 39 40 40 40 40 14 31 32 21
AAAAAAAAAAAAAAAAAAAAAAAAAAAAC 18 14 40 38 21 14 40 40 40 40 14 40 40 28 40 30 40 40 34 40 40 40 32 40 40 40 40 40 17 26 20 3
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 3 33 9 30 40 25 40 40 40 40 9 40 40 40 40 6 40 40 40 40 40 40 40 40 40 40 40 40 40 39 31 8 5
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCS 8 29 40 40 3 26 40 40 40 29 23 40 22 22 40 19 30 40 25 40 40 40 6 40 34 36 25 6 19 8 7 11 2
CCCCCCCCCCCCCCCCCCCCCCCCCCCCTT 11 25 25 39 23 19 28 39 40 12 17 40 35 13 40 9 40 40 32 40 33 40 39 37 40 40 40 4 40 11 11 5 4
CTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 3 13 40 40 40 20 21 40 25 40 40 6 40 5 40 40 15 40 40 32 40 40 40 11 40 39 40 21 40 21 17 8 12
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGR 11 33 17 40 34 19 40 40 40 40 40 15 40 40 40 40 39 9 38 40 12 40 37 40 40 31 40 40 40 40 40 19 6 14 3
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 8 39 38 34 14 27 19 40 40 40 26 29 40 40 40 40 40 22 4 27 40 40 40 40 40 40 31 40 40 14 36 4 24 9
CTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 19 11 40 40 32 22 40 40 40 40 40 14 40 40 40 40 28 40 40 40 40 40 40 40 40 40 40 40 26 40 24 23 29
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAM 19 22 26 14 9 35 40 40 40 24 35 40 40 40 36 33 32 17 40 40 33 40 40 40 40 34 25 40 40 39 15 40 33 2
GKTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 4 3 11 20 40 25 29 17 40 40 39 40 40 22 40 40 40 40 40 26 24 40 40 40 40 40 40 40 40 40 40 24 31 21
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC 10 24 21 12 18 12 14 40 40 40 17 29 40 40 29 29 40 9 11 27 40 40 33 40 40 20 40 40 40 32 40 3 22 9
CCCCCCCCCCCCCCCCCCCCCCCCCCCCMT 5 11 27 33 17 12 40 40 34 32 19 23 40 35 40 26 40 14 35 40 28 40 40 40 32 40 32 28 24 16 13 40 14 3 6
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAM 5 11 8 34 3 18 32 35 37 38 40 40 40 40 20 29 22 23 16 35 40 32 40 40 40 35 40 30 20 14 17 28 3 40 39 2

```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAH  6 17 6 5 4 40 29 33 40 22 28 22 28 40 23 12 31 40 3 21 18 36 20 30 26 40 36 33 32 38 22 16 24 25 6 40 40 1
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT  3 12 40 17 9 16 40 40 38 40 40 40 40 40 40 40 40 13 35 12 32 3 40 22 40 40 37 40 40 29 37 40 40 6 17 21 40 40
CTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT  3 6 6 19 31 8 8 27 35 35 40 21 40 40 24 19 33 38 7 40 16 22 6 40 28 40 26 34 40 37 28 32 13 4 21 31 20 17
AACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCMY  4 4 8 22 17 9 [illegible] 27 40 1 37 40 40 3 40 7 6 4 27 26 35 40 12 27 2[illegible] 34 21 22 35 4[illegible] 40 9 3 2
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC  [illegible] 18 3 8 [illegible] 20 29 40 11 40 40 7 2[illegible] 4 6 [illegible] 35 40 22 27 [illegible] 33 40 [illegible] 6 2 [illegible] 17 19 23
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAMMT  40 4 13 32 7 25 40 14 40 40 28 37 28 31 38 40 4 22 [illegible] 18 40 2[illegible] 40 27 37 20 [illegible] 31 31 23 [illegible] 7 4 [illegible] 40 2 2 2
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAMT  6 18 22 3 12 14 13 10 7 40 8 15 32 33 40 40 40 13 13 26 20 13 40 13 21 11 9 26 14 8 21 40 13 40 40 1 4
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACHMMT  4 10 12 40 22 16 8 9 38 12 23 18 9 39 6 19 26 27 26 15 19 23 10 19 15 33 30 22 30 4 40 32 4 1 1 2 20
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCY  7 20 13 4 8 24 19 14 31 6 18 30 10 3 30 3 40 8 12 8 19 22 40 16 40 17 30 19 21 21 35 3 5 11 21 3
AAAAAAAAAAAAAAAAAAAAACCCCCCCCCCCHMTYY      4 17 6 37 29 40 7 27 13 23 12 16 28 13 40 9 30 12 19 25 19 40 16 9 15 29 9 22 24 40 11 1 2 8 3 1
AAAAAAAAAAAAAAAAAAAAAAAGGGGGGGGGGGGG       3 23 14 5 12 16 6 11 40 13 26 19 12 31 19 20 21 13 10 17 29 7 11 10 20 11 40 4 17 6 8 15 17 40 40
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAACMMW  6 10 8 2 12 17 8 27 6 34 13 19 11 9 28 40 10 8 5 13 37 7 18 13 18 7 16 25 13 40 40 3 2 2 2
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAACMT  3 3 4 4 23 14 7 20 7 13 8 28 14 5 35 40 20 10 4 18 40 5 15 21 31 21 12 6 40 4 40 40 8 2 4
AAAAAAAAAAAAAAAGGGGGGGGGGGGGGGGGG      3 8 23 13 40 6 10 13 7 16 18 9 7 9 40 6 6 18 32 10 15 5 40 12 10 40 19 7 7 40 13 12 10 13
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGNS  7 16 12 3 17 14 7 20 10 6 23 25 4 16 40 10 40 19 13 27 9 18 10 6 22 12 11 4 18 6 25 1 1
AAAAAAAAAAAAAAAAAAAGGGGGGGGGGGGGG  6 10 4 10 8 17 6 29 18 4 13 22 5 8 11 4 30 15 6 7 7 3 10 40 4 11 8 10 11 9 9 7 40
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAHMT  13 12 20 10 4 13 6 23 23 12 20 6 27 17 13 26 16 4 18 6 7 13 8 6 8 8 14 38 40 1 1 3
AAAAAAAAAAAAACCCCCCCCCCCCCCCCTTY      6 6 11 40 6 18 12 15 12 10 5 12 40 7 11 7 21 6 34 6 21 14 28 13 17 29 40 6 6 10 3 2
AAAAAAAAAAAAAAAACCCCCCCCCCCCCGHRW      8 8 16 7 17 6 11 13 39 15 14 9 35 4 33 6 11 23 40 11 8 20 10 11 10 6 6 40 4 1 1 2
AAAAAAAAAAATTTTTTTTTTTTTTTTY      6 15 12 13 6 13 17 10 10 4 5 6 13 28 23 35 13 28 16 15 30 9 19 4 36 7 6 12 1
CCCCCCCCCCCCCCHTTTTTTTTTTTTY      13 6 18 2 21 3 10 14 2 34 14 23 6 24 1 20 28 13 6 4 24 10 23 19 13 11 14 2
CCCCCCCCCCTTTTTTTTTTTTTTTT  12 13 15 6 13 9 4 6 4 8 8 13 40 23 30 38 23 24 36 40 32 4 36 22 40 6
AGGGGGGGGGGGGGGKTTTTTTTTTT      7 4 8 26 6 21 30 3 23 40 8 34 8 28 7 2 32 22 23 20 14 28 18 24 16 14
BCCGGGGGGGGGGGGGGGGGGGGGGS      1 3 6 5 9 6 5 7 8 6 5 3 19 29 3 32 16 4 15 5 6 12 5 5 33 2
GGGGGGGGGTTTTTTTTTTTTTTTT  9 21 6 8 13 4 7 4 10 5 4 9 22 16 20 30 24 14 40 30 13 19 16 14 40 4
AAMTTTTTTTTTTTTTTTTTTTTT  17 9 1 16 32 23 16 33 13 22 7 22 18 7 15 40 10 14 14 15 23 11 15 17 26 5
STTTTTTTTTTTTTTTTTTTTTTTT  2 6 18 20 21 24 21 40 22 25 20 27 40 28 9 6 40 11 8 6 22 19 11 28 40 11
AAAAAAAAAAAAAAAAAAMMTTTTTT  4 7 3 8 25 5 14 8 40 40 17 22 18 7 9 12 40 3 1 1 20 19 6 22 19 25
AAAAAAMTTTTTTTTTTTTTTTW  6 9 6 6 3 3 2 6 26 27 6 23 39 5 32 30 40 8 40 15 20 22 2
NTTTTTTTTTTTTTTTTT  1 11 40 38 15 36 10 23 27 40 40 6 26 26 9 40 36 32
GGGGGGGKSTT  6 6 6 13 40 21 21 1 2 5 18
CGGGGGGGG  3 3 4 3 40 21 11 10 40
TTTTTTTTT  7 21 15 5 35 40 10 15 40
CCCCCYYYY  9 4 40 40 40 1 1 1 1
AAAAAAW  2 4 40 40 10 40 2
TTTTTT  18 13 40 40 26 40
AGGGGG  3 4 3 40 3 40
AAAAAA  17 2 40 40 15 40
CGGGGT  4 40 40 4 40 8
TTTT  40 40 5 40
AAAA  36 40 11 40
TTTT  40 40 26 40
TTTT  40 40 15 40
TTTT  40 40 13 40
CCCY  40 40 40 1
CCT  40 40 1
AA  40 40
Cluster 1 Stop 814
```

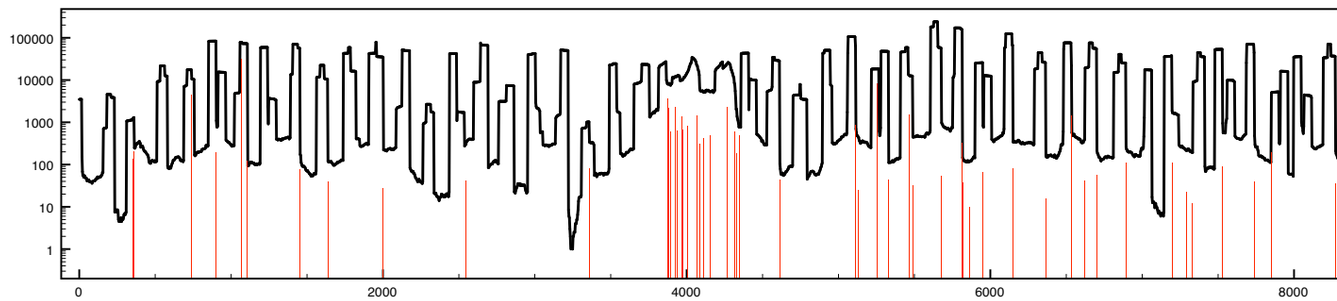
Analysis on a Shoestring - SNPs in Pooled Samples

- Tally base-by-base nucleotide frequencies and qualities
- Pick thresholds from simple error model
- 83% validated at predicted population level by RT-PCR!
- MAQ - frequent false positives and false negatives

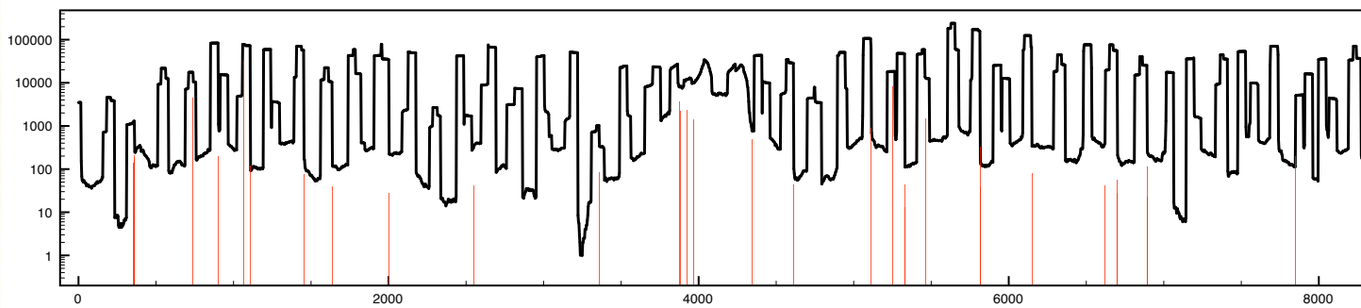
CONTIG	POSITION	DEPTH	VARIANT	V.DEPTH	%CALLS	%QUALS	V.QUAL	QUALS
scaffold_185	11915	62	G	59	0.9516	0.9891	1266	1280
scaffold_152	238099	90	T	85	0.9444	0.9221	1266	1373
scaffold_15	393751	77	A	75	0.9740	0.9945	1266	1273
scaffold_147	12004	60	G	50	0.8333	0.9806	1266	1291
scaffold_1133	6860	71	C	70	0.9859	0.9969	1266	1270
scaffold_52	103072	70	G	70	1.0000	1.0000	1265	1265
scaffold_3	680248	60	C	59	0.9833	0.9976	1265	1268
scaffold_229	75856	106	C	92	0.8679	0.9426	1265	1342
scaffold_213	71821	95	T	83	0.8737	0.9664	1265	1309
scaffold_15	391030	69	C	61	0.8841	0.9664	1265	1309
scaffold_140	135979	51	T	51	1.0000	1.0000	1265	1265
scaffold_14	953190	68	A	61	0.8971	0.9708	1265	1303
scaffold_58	76069	74	C	72	0.9730	0.9961	1264	1269
scaffold_49	374940	60	A	60	1.0000	1.0000	1264	1264
scaffold_35	1027666	53	G	51	0.9623	0.9929	1264	1273
scaffold_153	150	38	G	36	0.9474	0.9937	1264	1272
scaffold_152	131015	69	G	68	0.9855	0.9922	1264	1274
scaffold_13	611759	50	C	48	0.9600	0.9837	1264	1285
scaffold_1	3728402	50	G	47	0.9400	0.9821	1264	1287
scaffold_56	392560	50	C	49	0.9800	0.9875	1263	1279
scaffold_4739	644	74	C	73	0.9865	0.9945	1263	1270
scaffold_22	460616	53	T	53	1.0000	1.0000	1263	1263
scaffold_170	42834	71	A	64	0.9014	0.9730	1263	1298
scaffold_125	314020	59	G	51	0.8644	0.9776	1263	1292
scaffold_105	314291	68	C	64	0.9412	0.9914	1263	1274
scaffold_45	5453	72	C	70	0.9722	0.9961	1262	1267
scaffold_36	232766	82	T	76	0.9268	0.9875	1262	1278
scaffold_21	455896	78	A	71	0.9103	0.9813	1262	1286
scaffold_195	125468	65	C	47	0.7231	0.9575	1262	1318
scaffold_18	1400541	68	A	65	0.9559	0.9760	1262	1293
scaffold_177	197527	62	G	62	1.0000	1.0000	1262	1262
scaffold_176	59462	65	G	64	0.9846	0.9992	1262	1263
scaffold_167	112530	49	A	48	0.9796	0.9968	1262	1266



1% SNP ratio = 395 loci



>=5% SNP ratio = 74 loci



>= 10% SNP ratio = 42 loci

Analysis on a Shoestring - *De Novo* Transcriptome Assembly

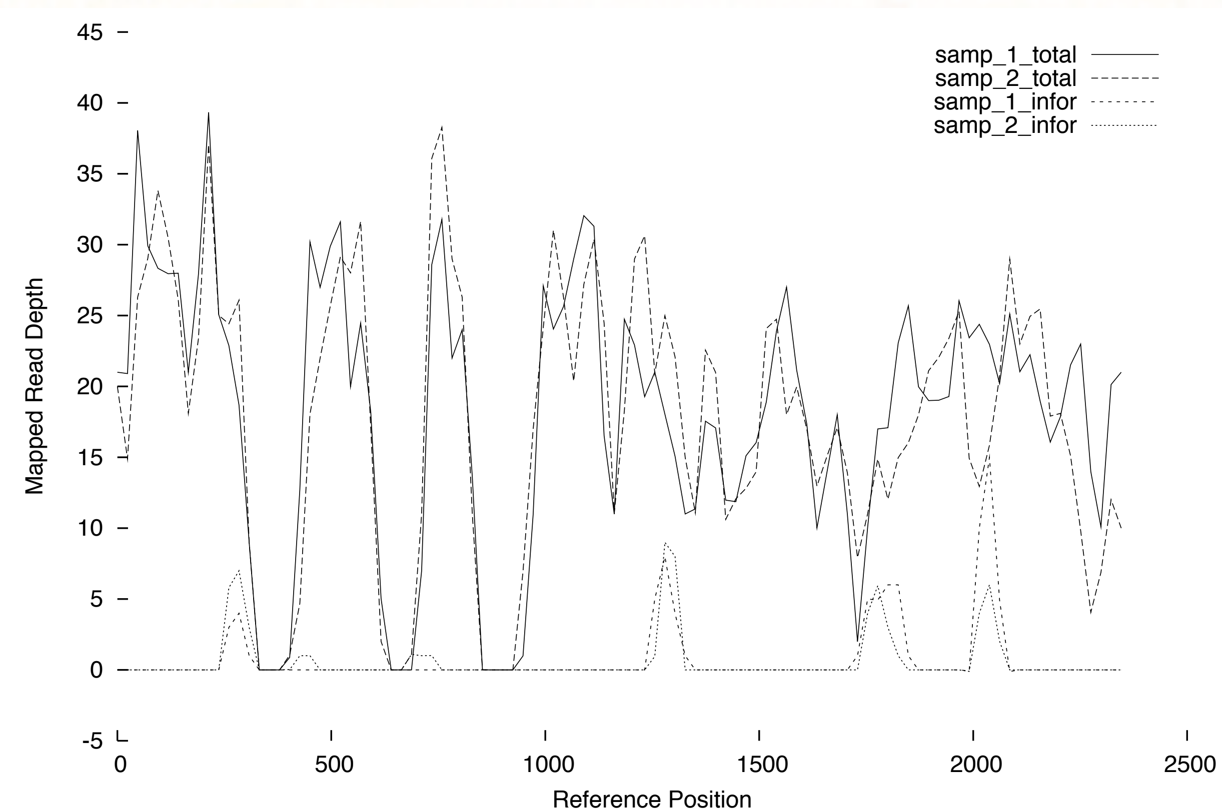
- Fix bugs in AMOScmp-shortRead … makes no joins after 3 days
- Try to install Euler-SR…failed & no one we know can get it installed either ;-)
- Add 2 new strategies to Velvet (default De Brujin graphs are too naïve)
  - Each increases mean/median contig length by up to 2X (4X total, constant total size)
- Simple alignments + graph theory to aggregate results of Multi-stage assemblies
- Align reads back to assembly for expression (& to avoid generating/parsing Velvet TXT files)
- HMMER3 alpha identifies all of the Core Eukaryotic Genes (CEGMA set of 437)

gene	167	0	length	1602	frame	5	Ribonucleotide reductase, alpha subunit
gene	149	0	length	2221	frame	3	N-arginine dibasic convertase NRD1 and related Zn
gene	115	5.20E-277	length	1550	frame	3	AAA+-type ATPase
gene	89	3.50E-276	length	1348	frame	1	DNA replication licensing factor, MCM5 component
gene	278	1.20E-261	length	1451	frame	4	U5 snRNP spliceosome subunit
gene	56	6.50E-256	length	2198	frame	6	RNA helicase
gene	102	3.90E-231	length	1263	frame	5	Phosphoglucomutase
gene	104	4.00E-228	length	1379	frame	4	WD40 repeat nucleolar protein Bop1, involved in rib
gene	169	1.10E-204	length	1192	frame	6	RNA polymerase II transcription initiation/nucleotide
gene	295	1.30E-190	length	924	frame	2	Predicted P-loop ATPase fused to an acetyltransfera
gene	88	4.90E-183	length	819	frame	4	DNA replication licensing factor, MCM2 component
gene	38	1.20E-179	length	980	frame	1	RNA polymerase III, large subunit
gene	188	2.30E-174	length	1378	frame	3	GDP-mannose pyrophosphorylase/mannose-1-pho
gene	31	1.10E-170	length	750	frame	3	Alanyl-tRNA synthetase
gene	159	4.00E-170	length	979	frame	1	Vesicle coat complex COPI, beta subunit
gene	182	7.10E-170	length	949	frame	5	Glucosamine 6-phosphate synthetases, contain am
gene	351	2.20E-169	length	1020	frame	6	Phenylalanyl-tRNA synthetase
gene	8	3.80E-168	length	1416	frame	4	ATPase component of ABC transporters with duplic
gene	83	2.80E-159	length	883	frame	5	Mitochondrial translation elongation factor Tu
gene	86	2.10E-158	length	720	frame	3	U5 snRNP-specific protein
gene	147	3.40E-156	length	686	frame	2	Nuclear exosomal RNA helicase MTR4, DEAD-box s
gene	47	1.50E-154	length	1155	frame	2	Ribosome Assembly protein
gene	90	1.00E-153	length	857	frame	3	HAT repeat protein
gene	176	1.40E-150	length	810	frame	1	Acyl-CoA synthetase
gene	180	4.90E-150	length	785	frame	5	Karyopherin (importin) beta 1
gene	68	2.20E-146	length	673	frame	1	Serine/threonine specific protein phosphatase involv
gene	153	3.30E-144	length	688	frame	3	Vesicle coat protein clathrin, heavy chain
gene	80	3.30E-144	length	615	frame	5	Isoleucyl-tRNA synthetase
gene	233	4.50E-144	length	642	frame	3	Ribonucleotide reductase, beta subunit
gene	148	7.70E-144	length	757	frame	4	DNA/RNA helicase MER3/SLH1, DEAD-box superfa
gene	229	1.20E-141	length	611	frame	1	26S proteasome regulatory complex, subunit RPN1
gene	46	5.20E-140	length	666	frame	3	Vesicle coat complex COPI, alpha subunit
gene	170	1.90E-136	length	578	frame	1	mRNA cleavage and polyadenylation factor II comp
gene	219	4.90E-135	length	637	frame	6	NADP-dependent isocitrate dehydrogenase
gene	69	1.10E-133	length	673	frame	6	Serine/threonine specific protein phosphatase involv
gene	168	1.10E-133	length	758	frame	1	RNA polymerase II transcription initiation/nucleotide
gene	140	4.50E-132	length	1077	frame	2	DEAH-box RNA helicase
gene	4	4.70E-131	length	736	frame	1	Structural maintenance of chromosome protein 1 (s
gene	320	1.70E-130	length	1023	frame	2	Phosphoglucomutase/phosphomannomutase
gene	232	1.10E-126	length	678	frame	4	Conserved protein Mo25
gene	145	3.10E-126	length	615	frame	1	Adaptor complexes medium subunit family
gene	33	2.20E-124	length	571	frame	4	P-type ATPase
gene	172	3.10E-124	length	555	frame	4	Glutaminyl-tRNA synthetase
gene	319	5.70E-122	length	512	frame	6	RNA polymerase II elongator complex, subunit ELP
gene	280	2.70E-121	length	685	frame	1	Ubiquitin fusion-degradation protein
gene	152	2.10E-119	length	555	frame	2	DNA polymerase delta, catalytic subunit
gene	114	3.10E-118	length	589	frame	6	AAA+-type ATPase containing the peptidase M41 d
gene	210	4.00E-118	length	587	frame	5	26S proteasome regulatory complex, subunit RPN6
gene	406	9.30E-118	length	673	frame	4	Predicted RNA-binding protein Pno1p interacting w
gene	363	4.20E-117	length	1352	frame	5	sn-1,2-diacylglycerol ethanolamine- and cholinepho
gene	122	1.00E-116	length	777	frame	2	Dehydrogenase kinase

COFACTOR GENOMICS

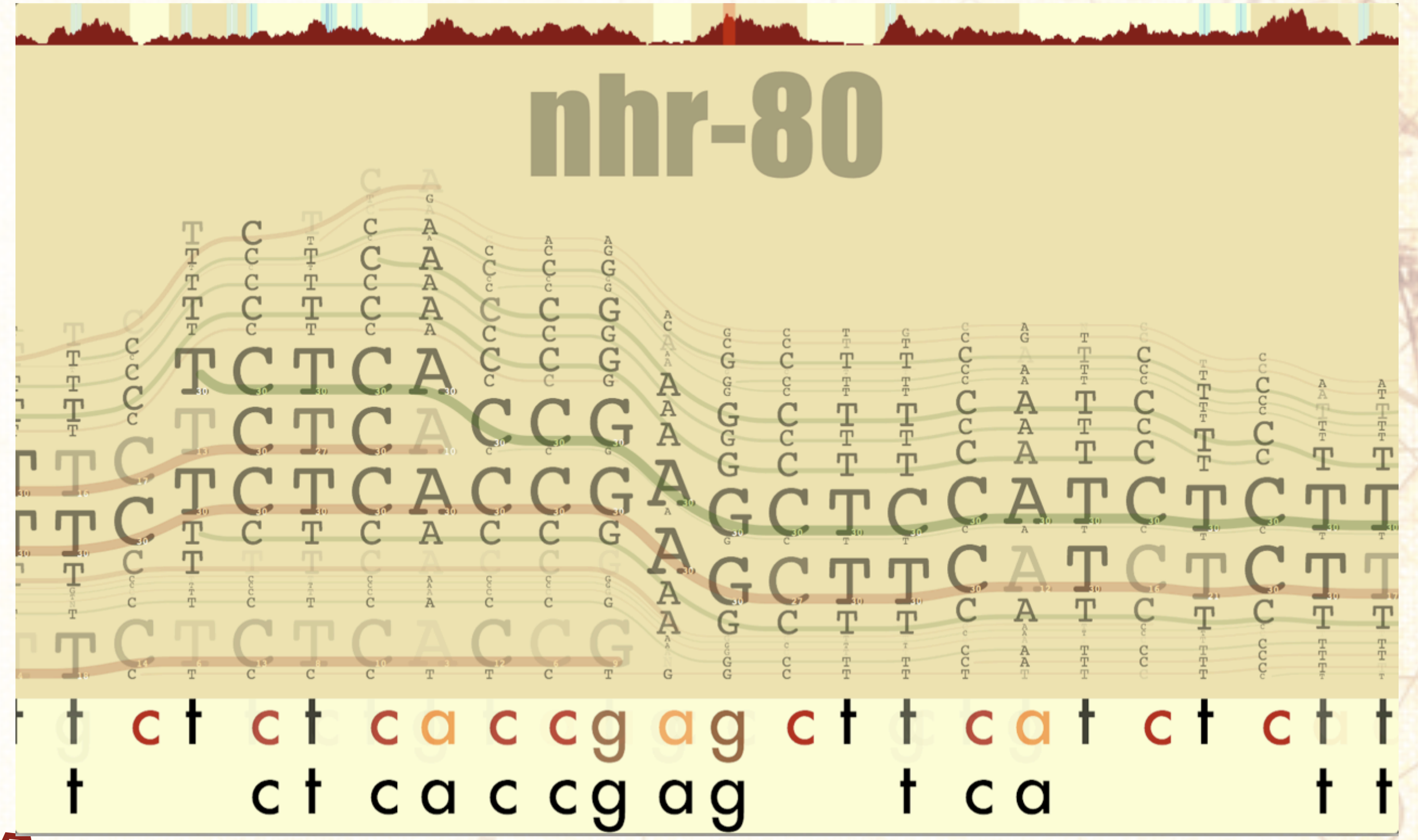
Analysis on a Shoestring - Allele Specific Expression

- Sequencing from F1 Generation of In-bred P's
- Competitive Mapping to P Genomes
- Compute fractional, parallel base-by-base coverage
- Look for genes with high coverage in only 1 genome




```
>My FASTA-Like Coverage File
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
0 0.23 0.23 0.35 0.57 0.57 0.57 0.78 0.78 1.10 1.10 1.23 1.23 1.23 1.23 1.23 1.33 1.33 1.33 1.33 1.49 1
.18 2.18 2.31 2.31 2.31 2.31 2.63 2.63 2.63 2.63 2.63 2.63 2.63 2.63 2.80 3.00 3.00 3.00 3.20 3.20 3.37
 3.48 3.35 3.35 3.35 3.32 3.32 2.99 2.99 2.99 2.99 3.16 3.28 3.28 3.18 3.18 3.18 3.45 3.28 3.08 2.98 2.
47 2.47 2.47 2.57 2.24 2.33 2.33 2.33 2.33 2.33 2.33 2.33 2.16 1.96 1.96 1.96 1.76 1.93 1.86 1.86 1.86
1.43 1.43 1.26 1.26 1.26 1.26 1.13 1.13 0.97 0.84 0.84 0.84 0.97 0.97 0.70 0.70 0.70 0.70 1.12 1.12 1.1
4 2.14 2.14 2.06 2.06 2.06 2.06 2.06 2.06 2.06 2.06 2.06 2.06 2.06 2.06 1.89 1.79 1.79 1.79 1.79 1.79 1
.67 1.67 1.67 1.67 1.67 1.67 1.67 1.67 1.67 1.67 1.54 1.54 1.54 1.54 1.54 1.54 1.12 1.12 1.12 0.59 0.39
 0.00 0.00 0.00 0.00 0.12 0.25 0.25 0.25 0.38 0.38 0.38 0.38 0.50 0.67 0.67 0.67 0.81 0.81 0.93 0.93 0.
06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.26 1.26 1.26
3.33 3.53 3.58 3.45 3.45 3.45 3.58 3.58 3.68 3.68 3.55 3.63 3.63 3.63 3.62 3.62 3.59 3.72 3.72 3.72 3.7
3 5.13 5.13 5.38 5.38 5.38 5.38 5.55 5.55 5.55 5.55 5.55 5.55 5.55 5.55 5.55 5.35 5.35 5.35 4.97 4.97 4
.21 3.21 3.21 3.21 2.96 2.96 2.96 2.96 2.96 2.71 2.71 2.71 2.67 2.67 2.69 2.57 2.57 2.57 2.57 2.40 2.40
 1.00 1.10 1.10 1.10 0.93 0.93 0.93 0.93 0.93 0.93 0.93 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 1.06 0.
76 0.76 0.76 0.76 0.66 0.66 0.66 0.66 0.66 0.66 0.57 0.57 0.45 0.45 0.45 0.45 0.45 0.45 0.45 0.35 0.35
0.41 0.41 0.41 0.41 0.41 0.41 0.41 0.41 0.41 0.28 0.28 0.28 0.45 0.45 0.45 0.65 0.65 0.65 0.85 0.85 0.8
0 1.60 1.60 1.60 1.60 1.85 2.22 2.22 2.22 2.47 2.56 2.56 2.81 2.81 2.81 2.81 2.81 2.81 2.81 2.81 2.81 2
.86 4.11 4.11 4.11 4.11 4.11 4.11 4.36 4.36 4.36 4.19 4.19 4.19 3.99 3.99 4.08 3.88 3.88 3.88 3.71 3.54
 3.12 3.12 2.88 2.50 2.50 2.50 2.25 2.17 2.17 1.92 1.92 1.92 1.92 1.92 1.92 1.92 1.92 1.92 1.92 1.67 1.
33 0.33 0.33 0.33 0.33 0.08 0.08 0.08 0.08 0.08 0.08 0.08 0.08 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0
.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
```

Cofactor Browser



COFACTOR GENOMICS

Multi-Faceted Sequencing & Analysis

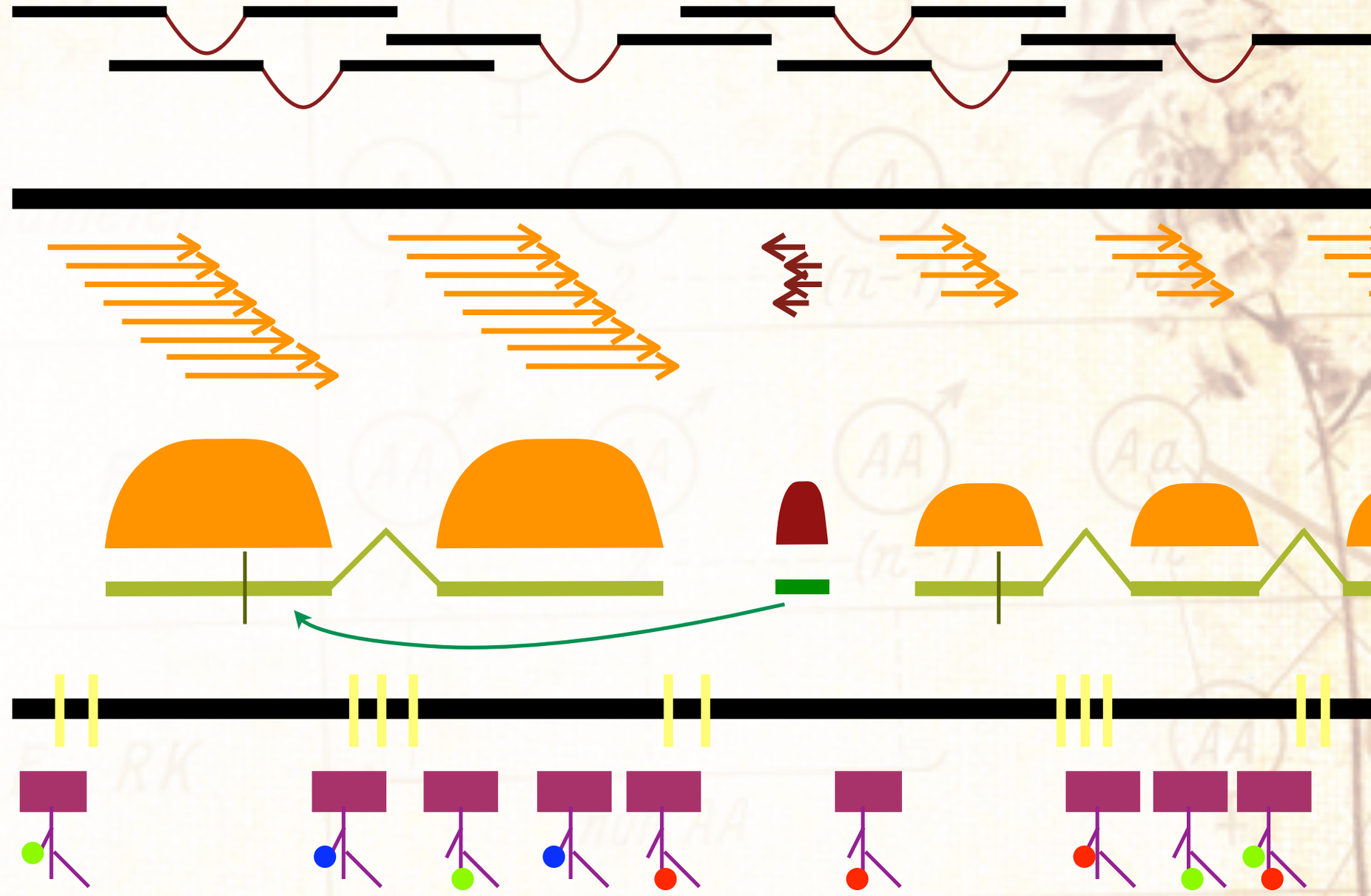
Assembly

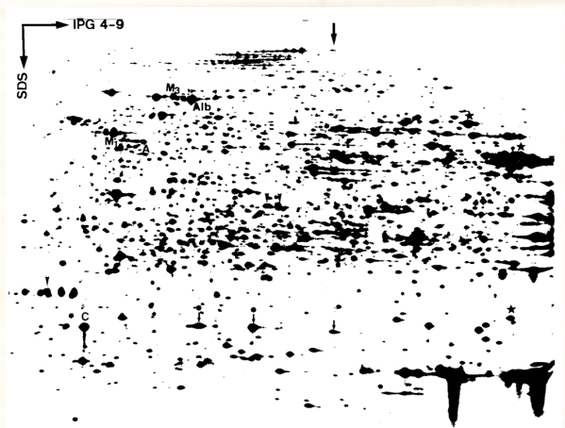
Expression Analysis

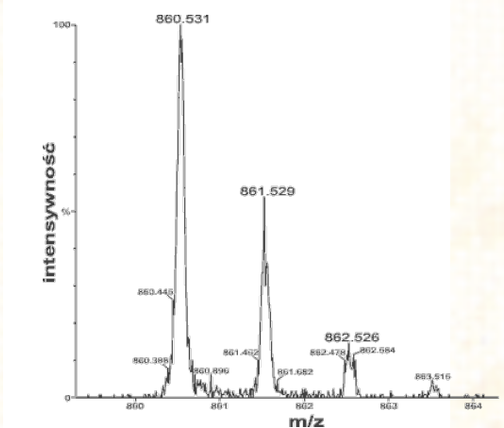
Gene Annotation

Nucleosome Profiling

Mass Spec & 2D PAGE

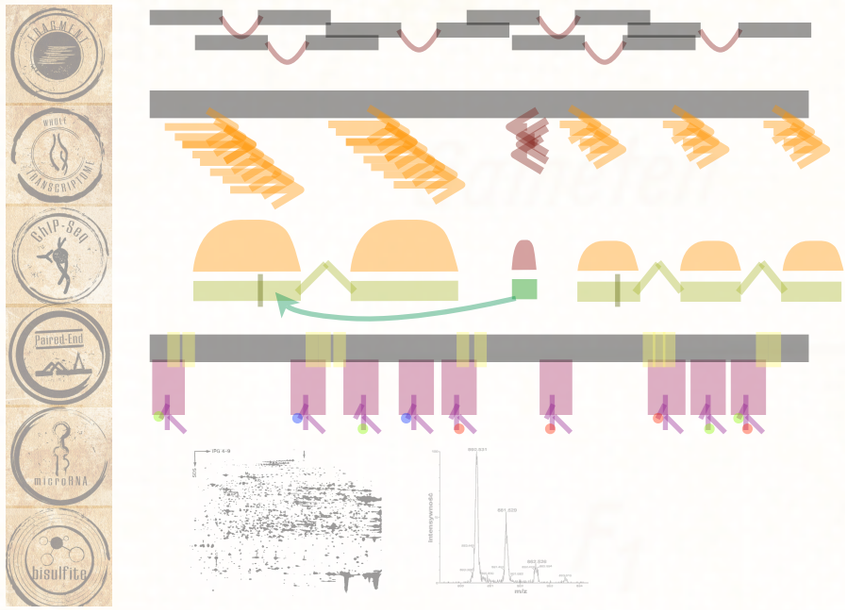
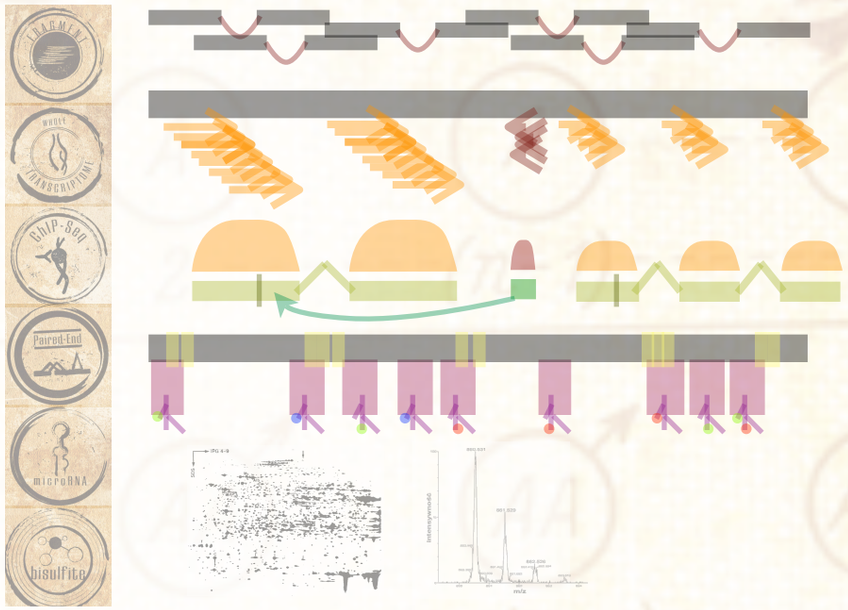
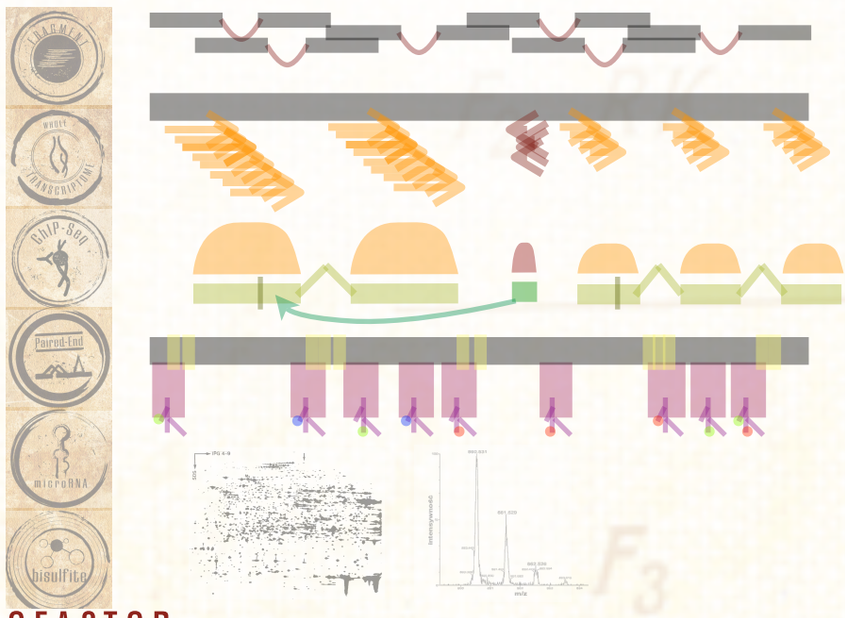
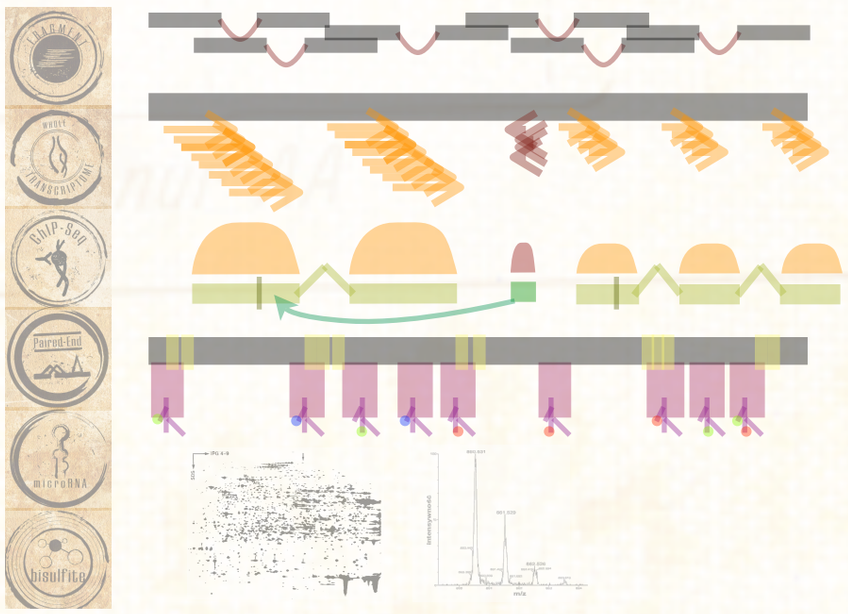






COFACTOR GENOMICS

Even Higher Order: Temporal/Spatial

	Time Point 1	...	Time Point X
Tissue 1		...	
...	...	...	...
Tissue Y		...	

Highest Order Data & Computation

Public DB

Project 1

...

Project N

Investigator 1 ... Investigator N

How Can We Get There Analysis-wise?

- No More Text files
- No More Perl Engines
- No More Single-threaded Apps
- No More Clusters, Configuration, or Installation
- Standard Indexed Binary Formats
- Free, Universal Computational Engine
- Open & Extensible System for Everyone

COFACTOR GENOMICS

Our Solution - The Genome Operating System

GUI

PDF

NovoAlign

PRB

Expression

Applications

Gentuition UI

'CPU'

'File System'

PostgeSQL

Genomes

Reads

Annotations

Alignments

Programming API

SNPs

Meta Data

Network

COFACTOR GENOMICS

Genome OS - Design Principles

- No Cluster Required **Efficient Algos, Fancy structs, 100% Multithreaded**
- No Petabyte Disks Required **Custom Binary Storage & Intermediates**
- Zero-Configuration
- Platform Independent **All Java 'Binaries'**
- Scales to Size of Project **Small footprint when sparse, low overhead when dense**
- Open & Extensible **Open specification & stellar API**
- Smart & Preemptive Computation **Meta-Data you actually care about**

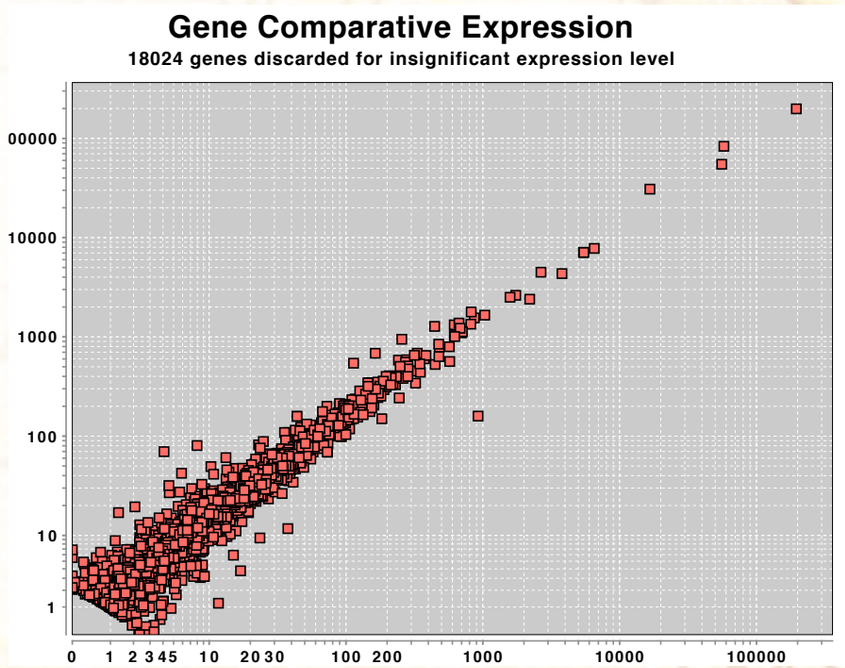
COFACTOR GENOMICS

Genome OS - Progress/Current Example

gentuition new named CancerProject
gentuition CancerProject set ProjectType RNAseq
gentuition CancerProject add sample named Tumor
gentuition CancerProject add sample named Normal
gentuition CancerProject add references from Human37.fasta
gentuition CancerProject/Tumor add alignments from tumor_novos/
gentuition CancerProject/Normal add alignments from normal_novos/
gentuition CancerProject Compare

In only 36 MINUTES:

- On one 8-core Apple Xserve, 32 GB RAM, 1TB Disk
- Parses 8 giga-bases of Paired-End alignments (160 GB)
- Organizes isoforms of all genes
- Computes gene-by-gene expression profile both samples
- Computes unannotated expression patches and depths
- Creates PDF graph of expression
- Stores it all back in 340 MB (0.2% of input)



COFACTOR GENOMICS

Genome OS - Time/Space Complexity

O(1)	To select genome "partition" that easily fits into RAM
O(Log M)	To seek to desired start location, M = Min(partition length, data entries)
O(1)	To seek to successor of any location
O(1)	Access to read, alignment, gene/transcript/exon, cluster, reference from any position
O(1)	Access to custom defined fields
Amortized O(1)	To keep search structures balanced
No Locking	Absolutely no locks or spinning, not even to rebalance search structures
All Single-Pass	Consensus, expression levels, novel read clusters, all computed as data is read from disk
Load Balanced	Parallelized **within** chromosomes and **within** files to be free from balance assumptions
No Re-Parsing	Data structures are stored as serialized Java objects
All Binary	All data is stored in the smallest possible primitive type and then compressed
Amortized O(1)	Space for search structure overhead, scales with lesser of genome or data size

COFACTOR GENOMICS

Genome OS - API Example

```
public class ExonCoverageTask extends AWGAAlignmentTask {

  ExonRegion exon = null;
  int exonTotalCoverage = 0;

  public ExonCoverageTask(){
     super(true);
  }

  public void setup(ExonRegion exon){
     this.exon = exon;
     this.exonTotalCoverage = 0;
  }

  public void executeTask(Alignment aln){
     int alnEnd = aln.alignmentLeftPosition + gentuition.readLength - 1;
     exonTotalCoverage += Math.max(Math.min(alnEnd, exon.getEnd())
                              - Math.max(aln.alignmentLeftPosition, exon.getStart()) + 1, 0);
  }

   public void cleanup(){
      this.exon.define("total-coverage", exonTotalCoverage);
   }
```

COFACTOR GENOMICS

Genome OS - API Example

```
public class ExonCoverageTask extends AWGAAlignmentTask {

  ExonRegion exon = null;
  int exonTotalCoverage = 0;

  public ExonCoverageTask(){
     super(true);
  }

  public void setup(ExonRegion exon){
     this.exon = exon;
     this.exonTotalCoverage = 0;
  }

  public void executeTask(Alignment aln){
     int alnEnd = aln.alignmentLeftPosition + gentuition.readLength - aln.getInsertions().size() + aln.getDeletions().size() - 1;
     exonTotalCoverage += Math.max(Math.min(alnEnd, exon.getEnd()) - Math.max(aln.alignmentLeftPosition, exon.getStart()) + 1, 0);

     for( Byte next : aln.getDeletions() ){
            int deleteLocation = aln.alignmentLeftPosition + next - 1;
            if(deleteLocation >= exon.getStart() && deleteLocation <= exon.getEnd()){
                  exonTotalCoverage--;
            }
      }
   }

   public void cleanup(){
       this.exon.define("total-coverage", exonTotalCoverage);
   }
}
```

COFACTOR GENOMICS

Genome OS - Current Development

GUI

PDF

MapReads

CSfasta

NovoAlign

PRB

Expression

Applications

Gentuition UI

'CPU'

'File System'

PostgeSQL

Genomes

Reads

Annotations

Alignments

Programming API

SNPs

Meta Data

Network

COFACTOR GENOMICS

Genome OS - Release Timeline

'CPU'

Programming API

Free & Open Spec

Community Web Site

September 1

NovoAlign

Ensembl GTF

'File System'

PostgeSQL

PRB

S/BAM

CSfasta

MapReads

Command Line

Meta Data

Open Source

December 1

GUI Display Protocol

Network

FS

Open Source

COFACTOR GENOMICS

Special Thank You
to my Computer Science Interns:

Michael Fahey

Jonathan Wald