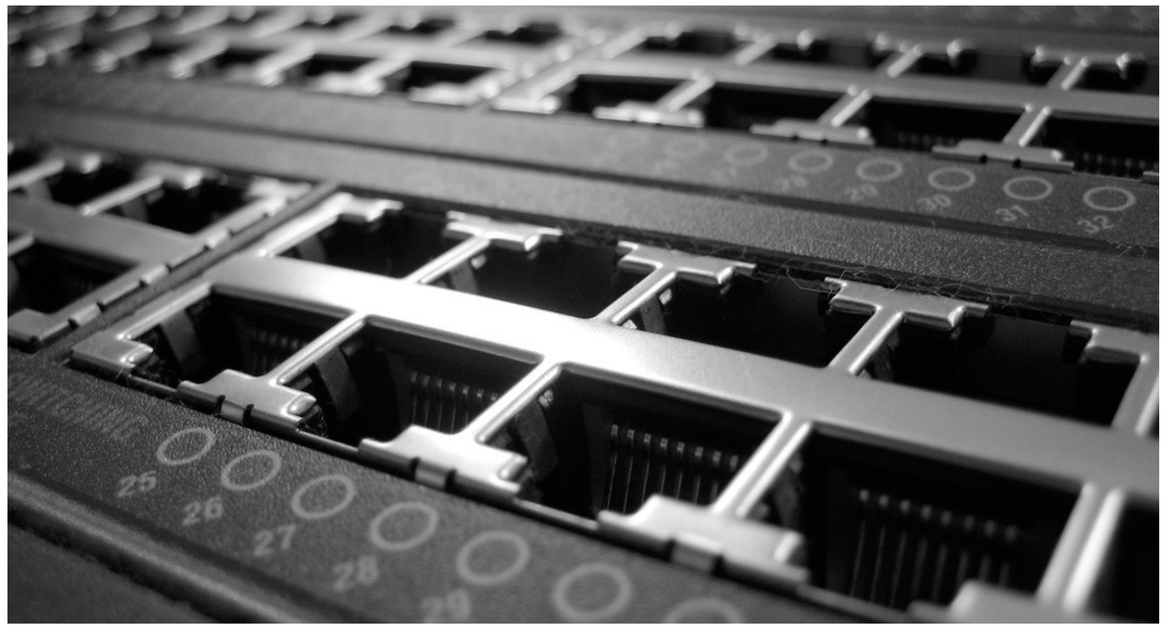
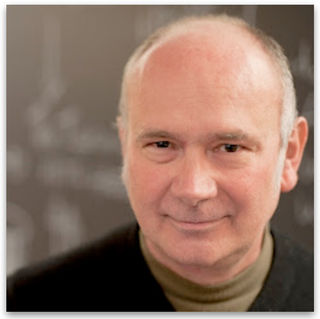


Translational bioinformatics and NGS

Center for Health Bioinformatics, Harvard School of Public Health

<intro>



Win Hide

Oliver Hofmann

Shannan Ho Sui

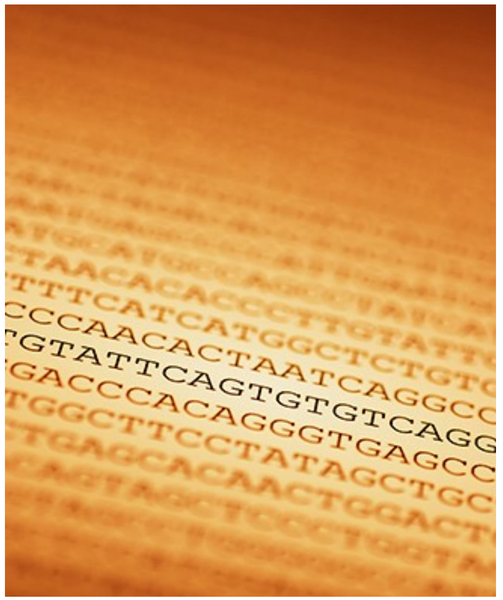
John Hutchinson

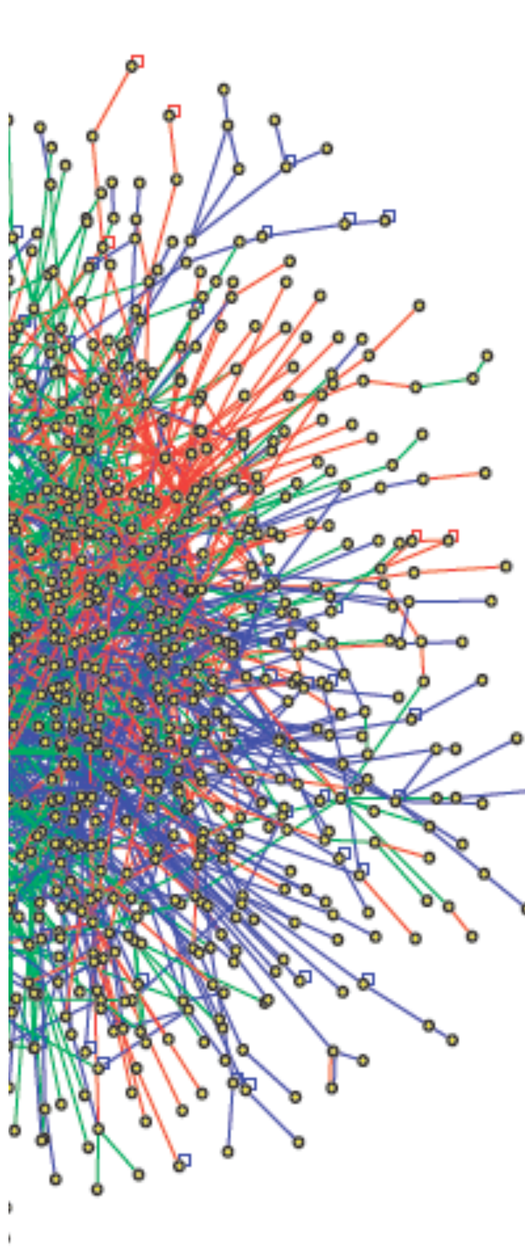
Meeta Mistry

John Morrissey

Rory Kirchner

Brad Chapman





Three focus areas

Research computing

Next-gen sequencing

Functional significance

HARVARD
SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

NIEHS / CFAR Bioinformatics Core

Center for Stem Cell Bioinformatics

Harvard Catalyst Bioinformatics Consulting

more than 300 consults from almost all Harvard-affiliated institutions

HARVARD
SCHOOL OF PUBLIC HEALTH

HARVARD STEM CELL INSTITUTE

HARVARD CATALYST
THE HARVARD CLINIC AND TRANSLATIONAL SCIENCE CENTER

HARVARD MEDICAL SCHOOL
Tools and Technology

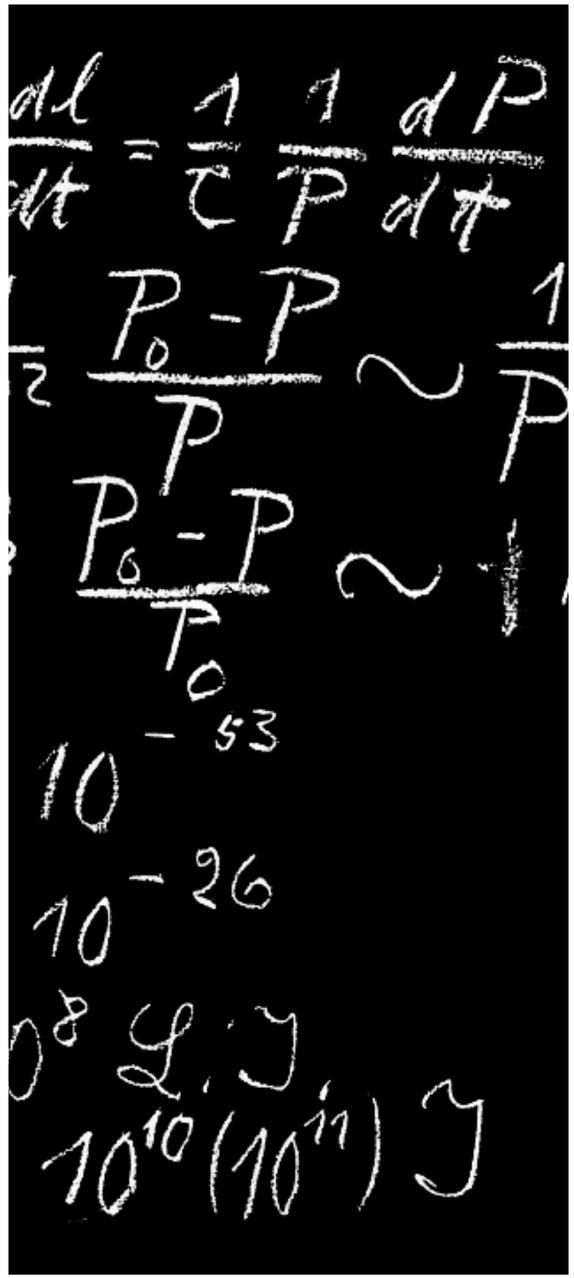
NIEHS / CFAR Bioinformatics Core

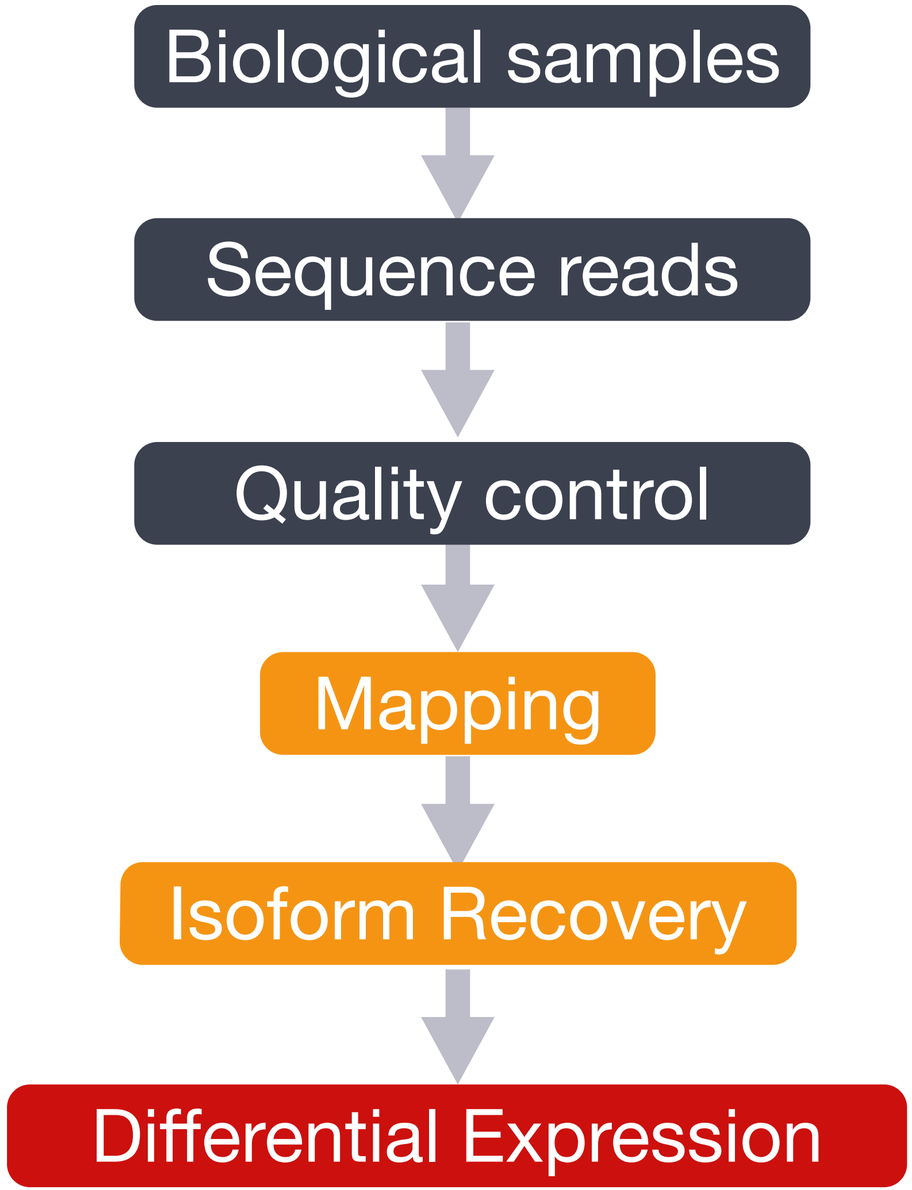
Center for Stem Cell Bioinformatics

Harvard Catalyst Bioinformatics Consulting

more than 300 consults from almost all Harvard-affiliated institutions

Scope


$\frac{dl}{dt} = \frac{1}{\tau}\frac{1}{P}\frac{dP}{dt}$
$\frac{P_0 - P}{P} \sim \frac{1}{P}$
$\frac{P_0 - P}{P_0} \sim 1$
10^{-53}
10^{-26}
$10^{10}(10^{11})$

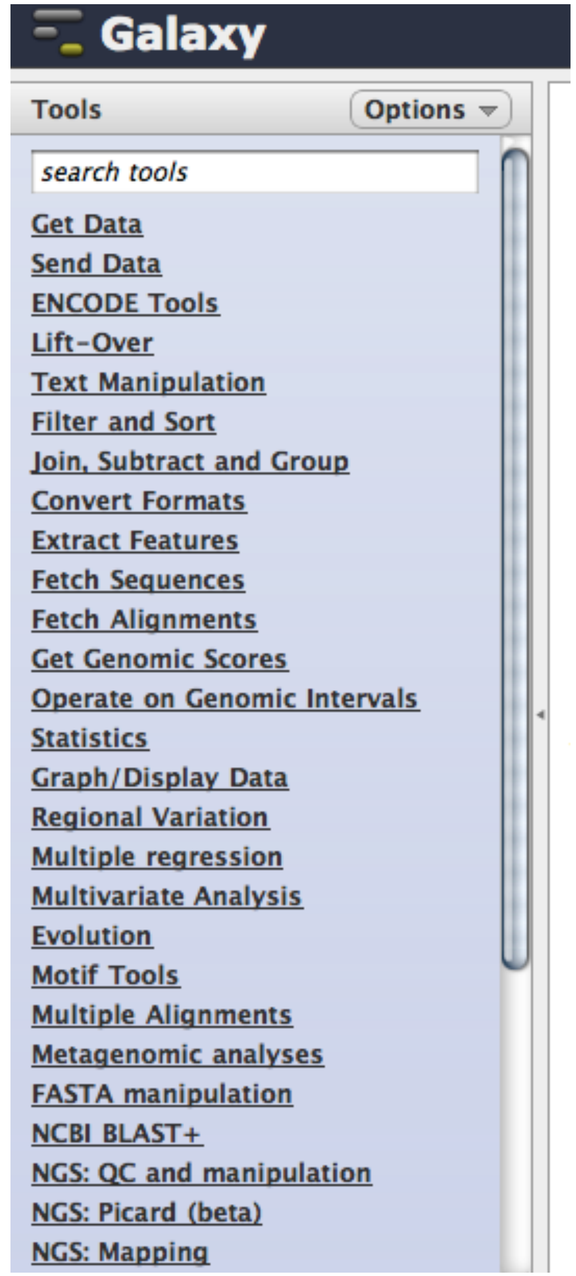

Biological samples
Sequence reads
Quality control
Mapping
Isoform Recovery
Differential Expression

Workflows

Using Galaxy

EC2 lab environment

http://23.23.134.25/



Do-it yourself

No black box

More control

Better understanding of the experiment

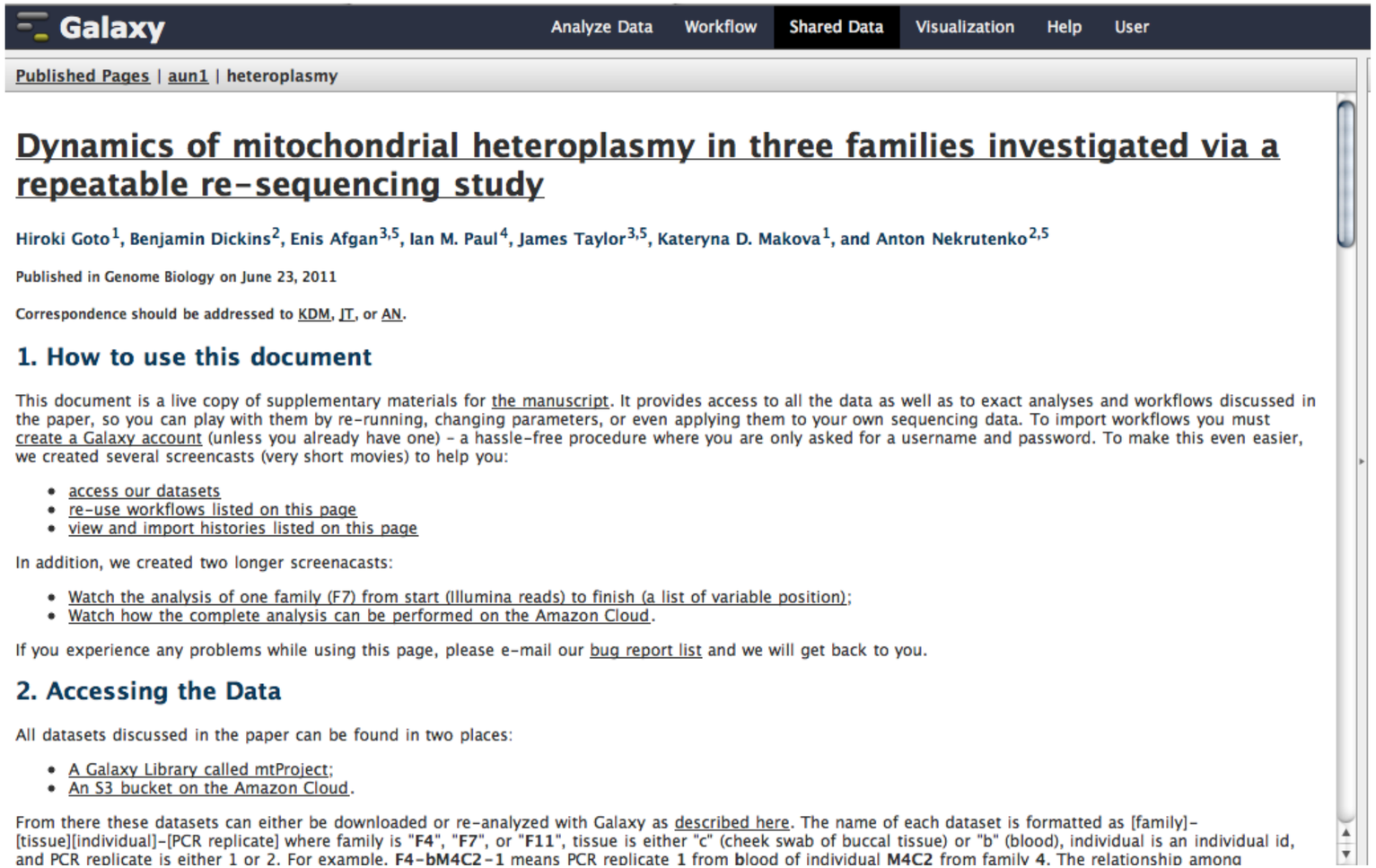
No command line / UNIX experience required

Galaxy Screencasts: The best way to understand how...

Online tutorials and screencasts

http://usegalaxy.org

http://galaxyproject.org


Galaxy

Analyze Data | Workflow | Shared Data | Visualization | Help | User

Published Pages | aun1 | heteroplasmy

Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study

Hiroki Goto[1], Benjamin Dickins[2], Enis Afgan[3,5], Ian M. Paul[4], James Taylor[3,5], Kateryna D. Makova[1], and Anton Nekrutenko[2,5]

Published in Genome Biology on June 23, 2011

Correspondence should be addressed to KDM, JT, or AN.

1. How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to all the data as well as to exact analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own sequencing data. To import workflows you must create a Galaxy account (unless you already have one) - a hassle-free procedure where you are only asked for a username and password. To make this even easier, we created several screencasts (very short movies) to help you:

- access our datasets
- re-use workflows listed on this page
- view and import histories listed on this page

In addition, we created two longer screenacasts:

- Watch the analysis of one family (F7) from start (Illumina reads) to finish (a list of variable position);
- Watch how the complete analysis can be performed on the Amazon Cloud.

If you experience any problems while using this page, please e-mail our bug report list and we will get back to you.

2. Accessing the Data

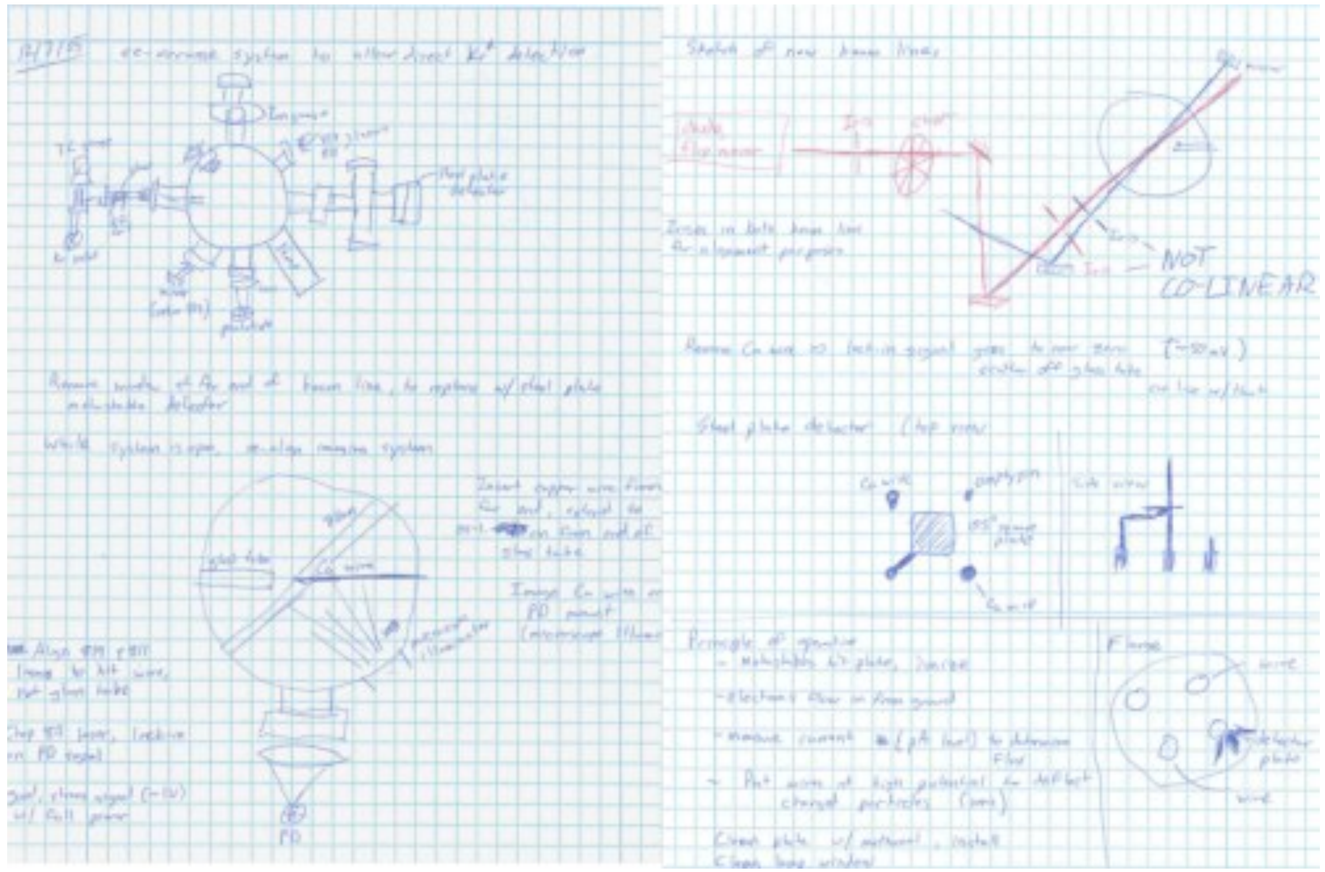
All datasets discussed in the paper can be found in two places:

- A Galaxy Library called mtProject;
- An S3 bucket on the Amazon Cloud.

From there these datasets can either be downloaded or re-analyzed with Galaxy as described here. The name of each dataset is formatted as [family]-[tissue][individual]-[PCR replicate] where family is "**F4**", "**F7**", or "**F11**", tissue is either "c" (cheek swab of buccal tissue) or "b" (blood), individual is an individual id, and PCR replicate is either 1 or 2. For example, **F4-bM4C2-1** means PCR replicate **1** from blood of individual **M4C2** from family **4**. The relationship among

Reproducible research

http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy



Reproducible research

Keep a lab book

ANALYSIS

nature genetics

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

Reproducible research

Don’t be afraid

Test... 2... 3

Course in pilot phase

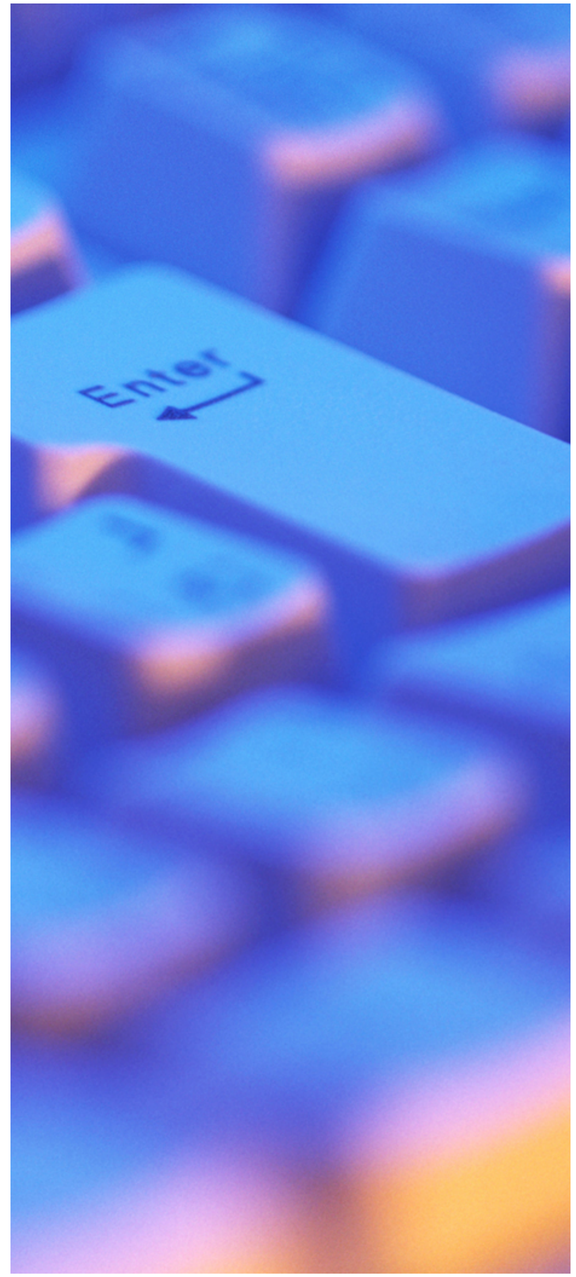
Collaborate

Feedback

- Survey
- Minute Cards

More information

See the course website for pointers



Contact

ohofmann@hsph.harvard.edu

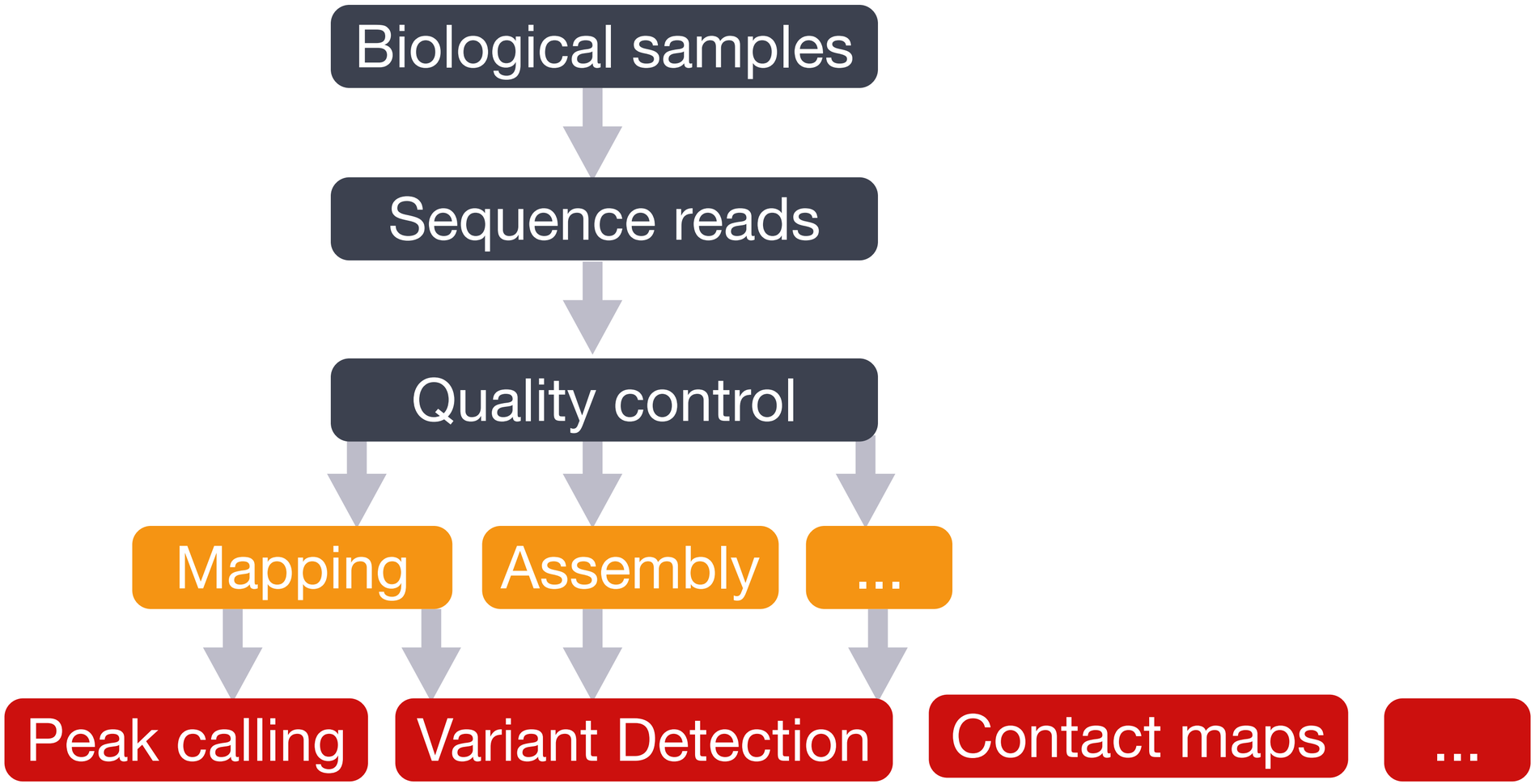
@fiamh

</intro>

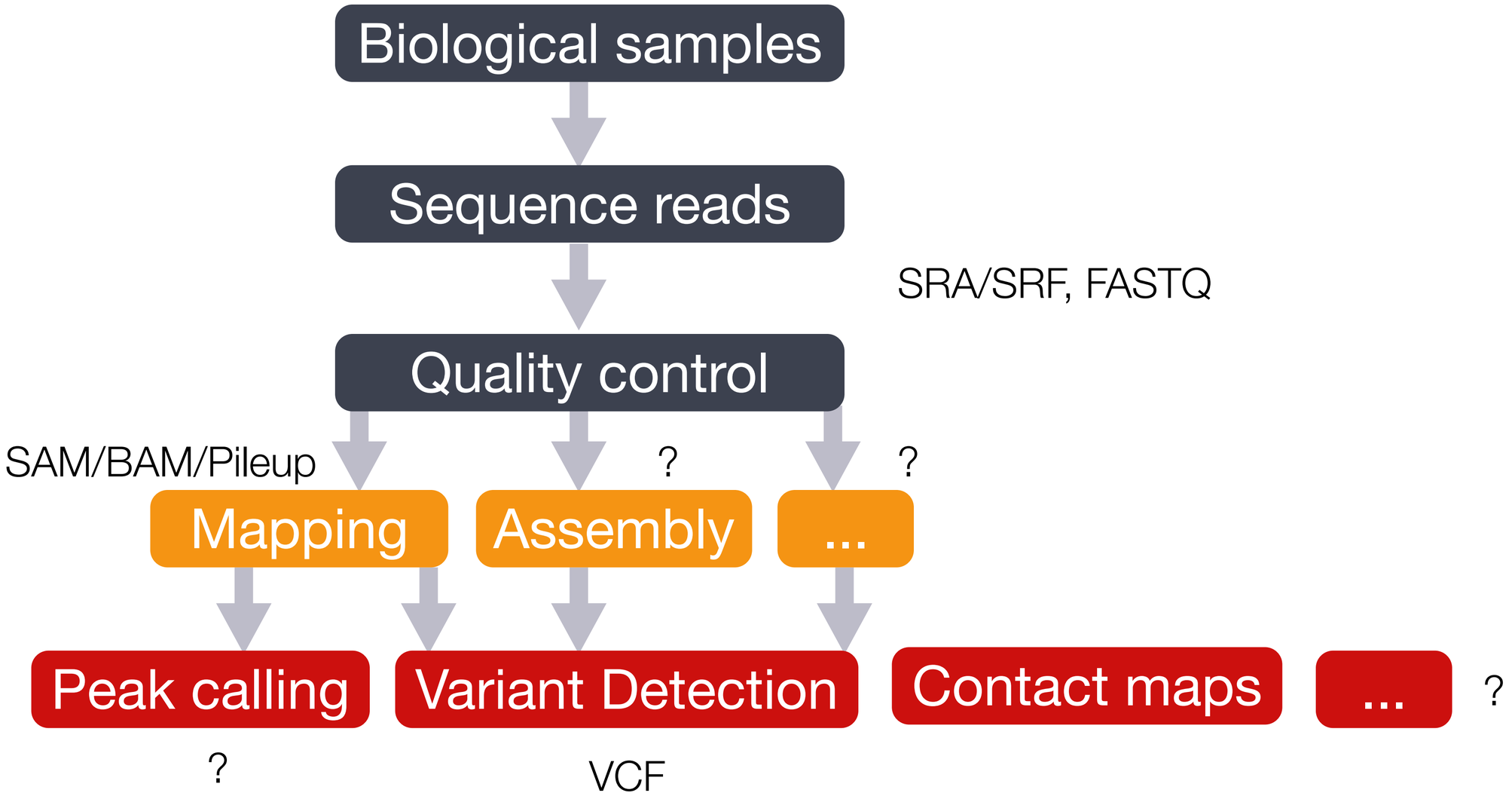
Need for standards

Plug and play: modular approach to tools

Vital factor in application acceptance


Biological samples
Sequence reads
Quality control
Mapping
Assembly
...
Peak calling
Variant Detection
Contact maps
...

Standards

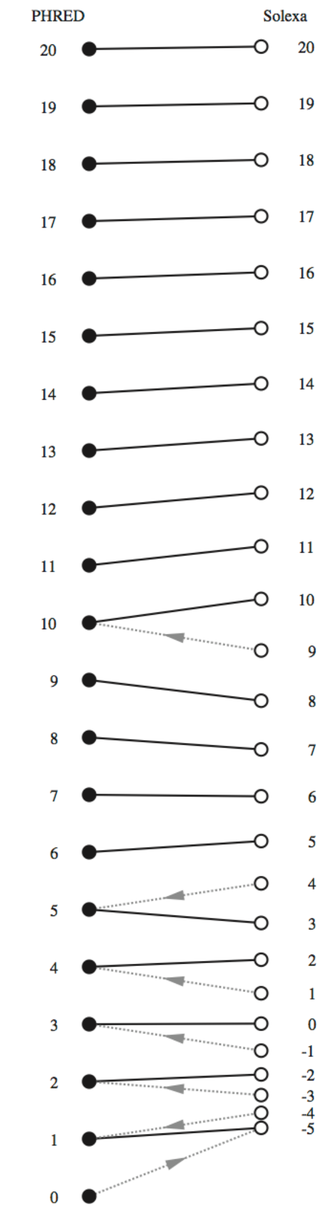

Biological samples
Sequence reads
SRA/SRF, FASTQ
Quality control
SAM/BAM/Pileup
?
?
Mapping
Assembly
...
Peak calling
Variant Detection
Contact maps
...
?
?
VCF

Standards

FASTQ: a “standard”

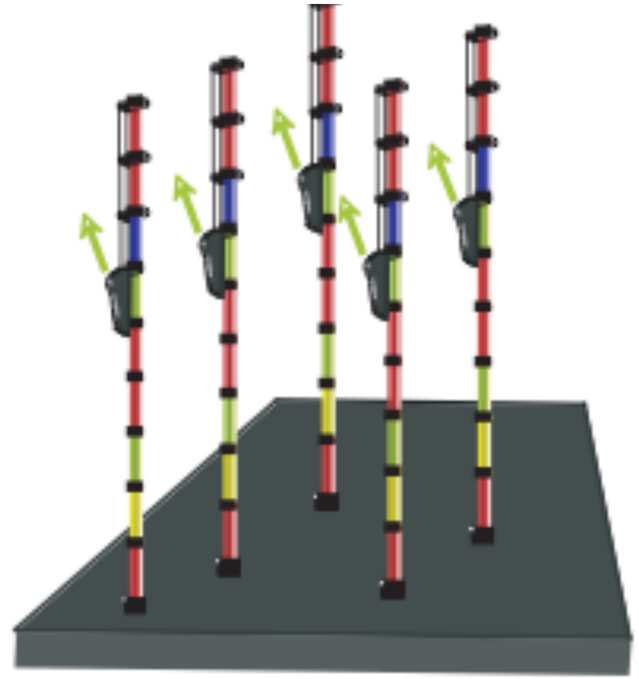
Sanger FASTQ, Solexa FASTQ, ABI Colour Space FASTQ, ...

@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93 3+&**$#""""""""""""""7F@71,’";C?,B;?6B;:EA1EA**
1EA5’9B:?:#9EA0D@2EA5’:>5?:%A;A8A;?9B;D@ /=<?7=9<2A8==

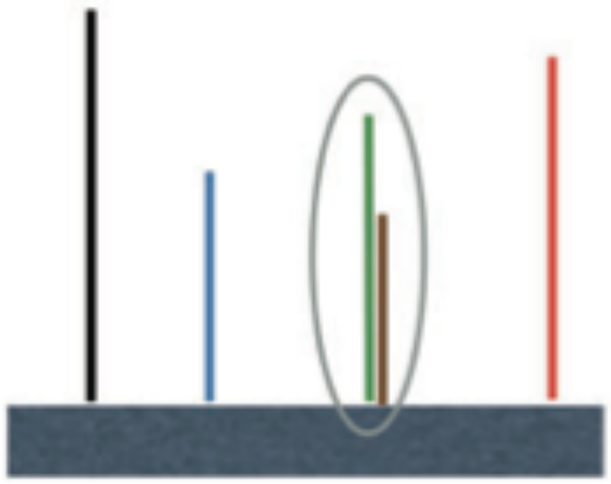


Error profiles

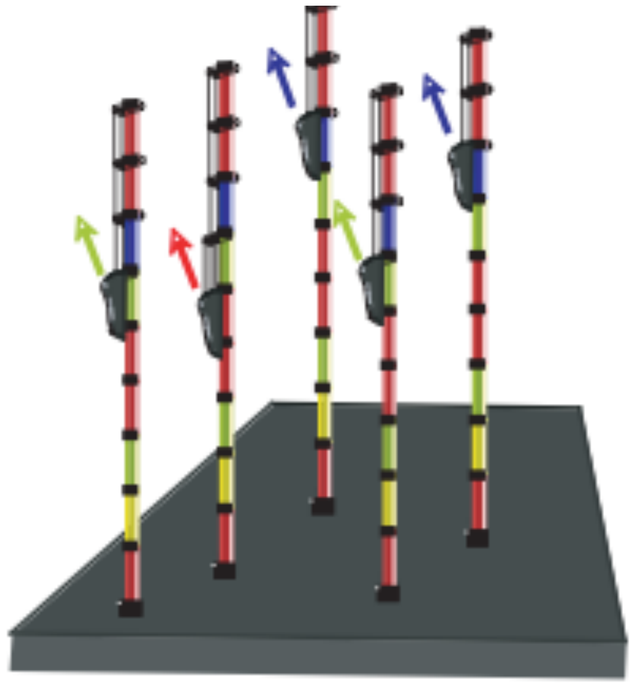
- PCR artifacts
- Error dependency on technology

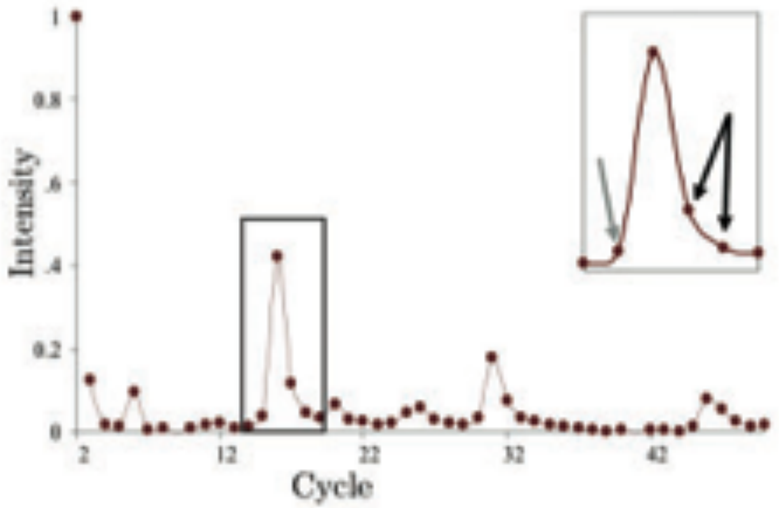


Illumina

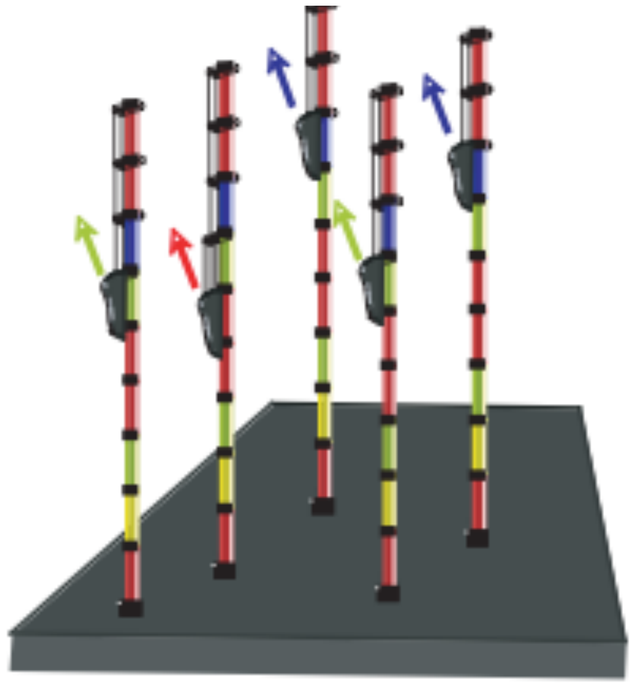


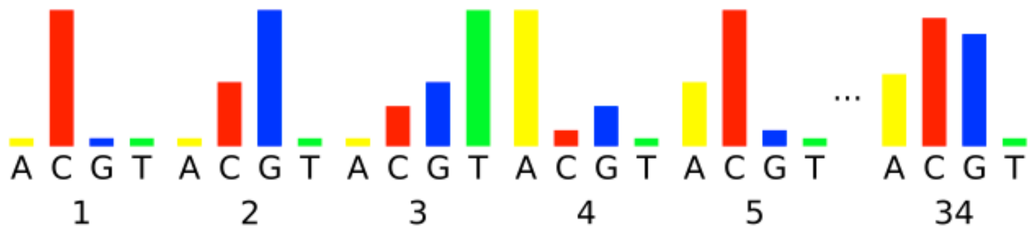
Illumina: mixed clusters



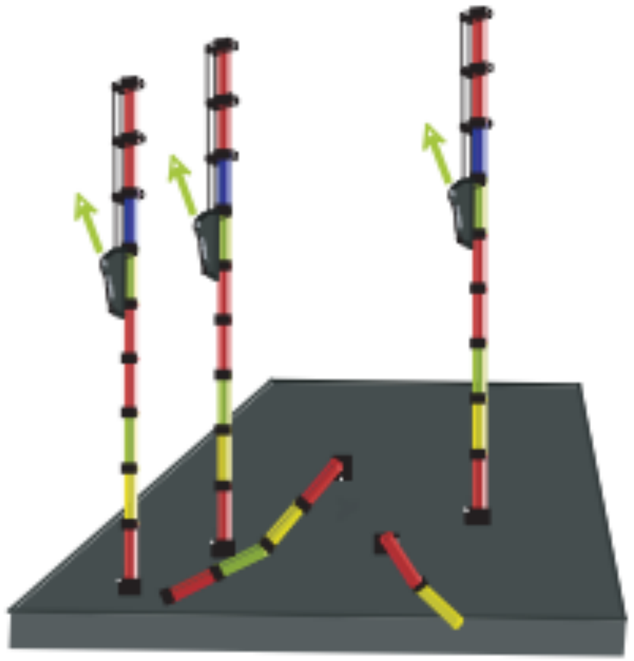


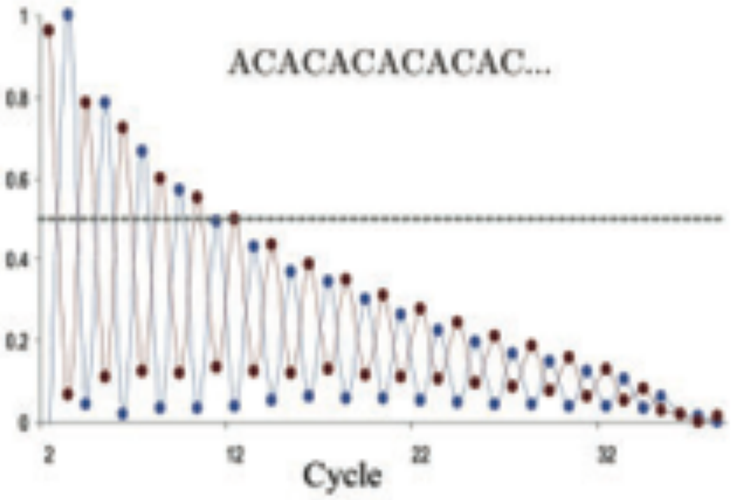
Illumina: phasing



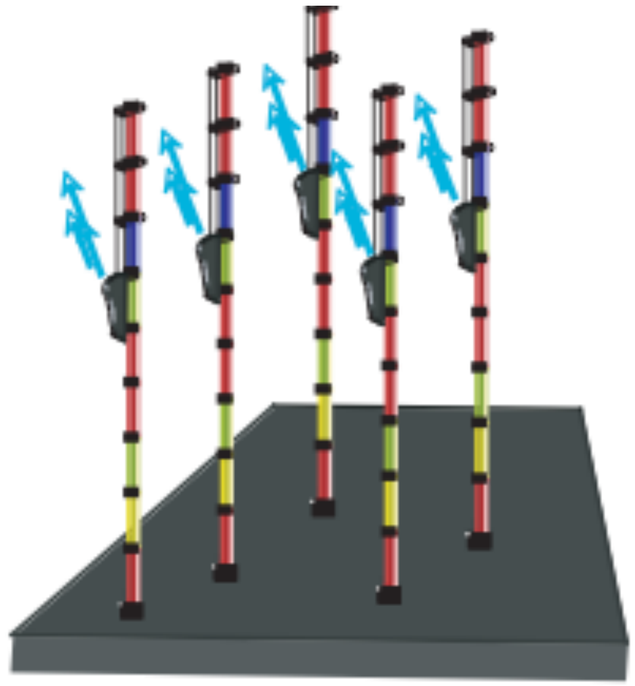


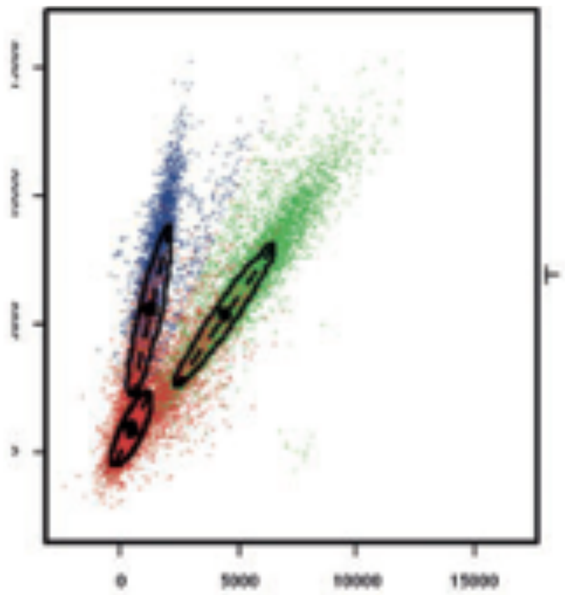
Illumina: phasing



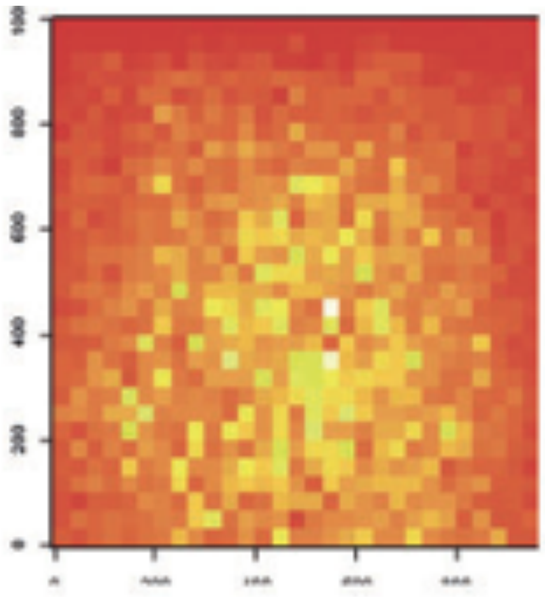


Illumina: signal decay

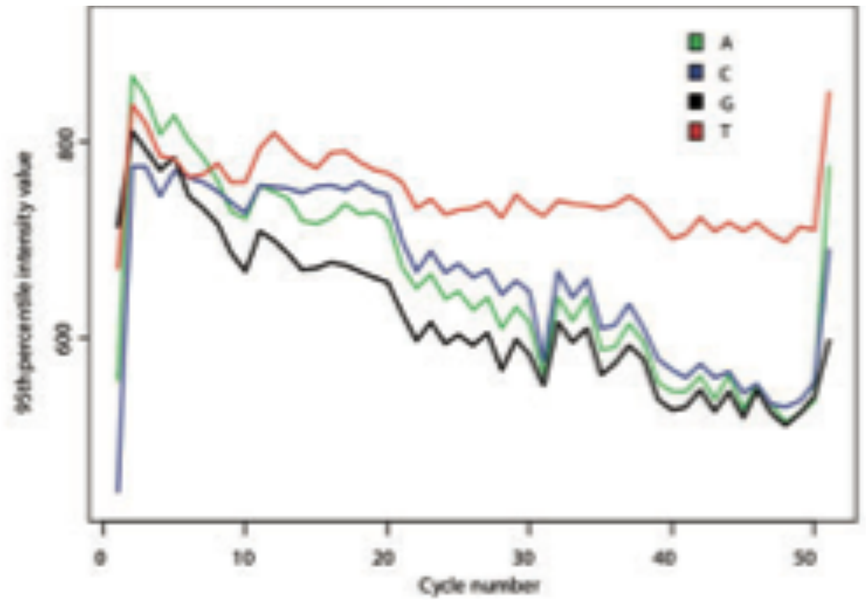




Illumina: cross-talk

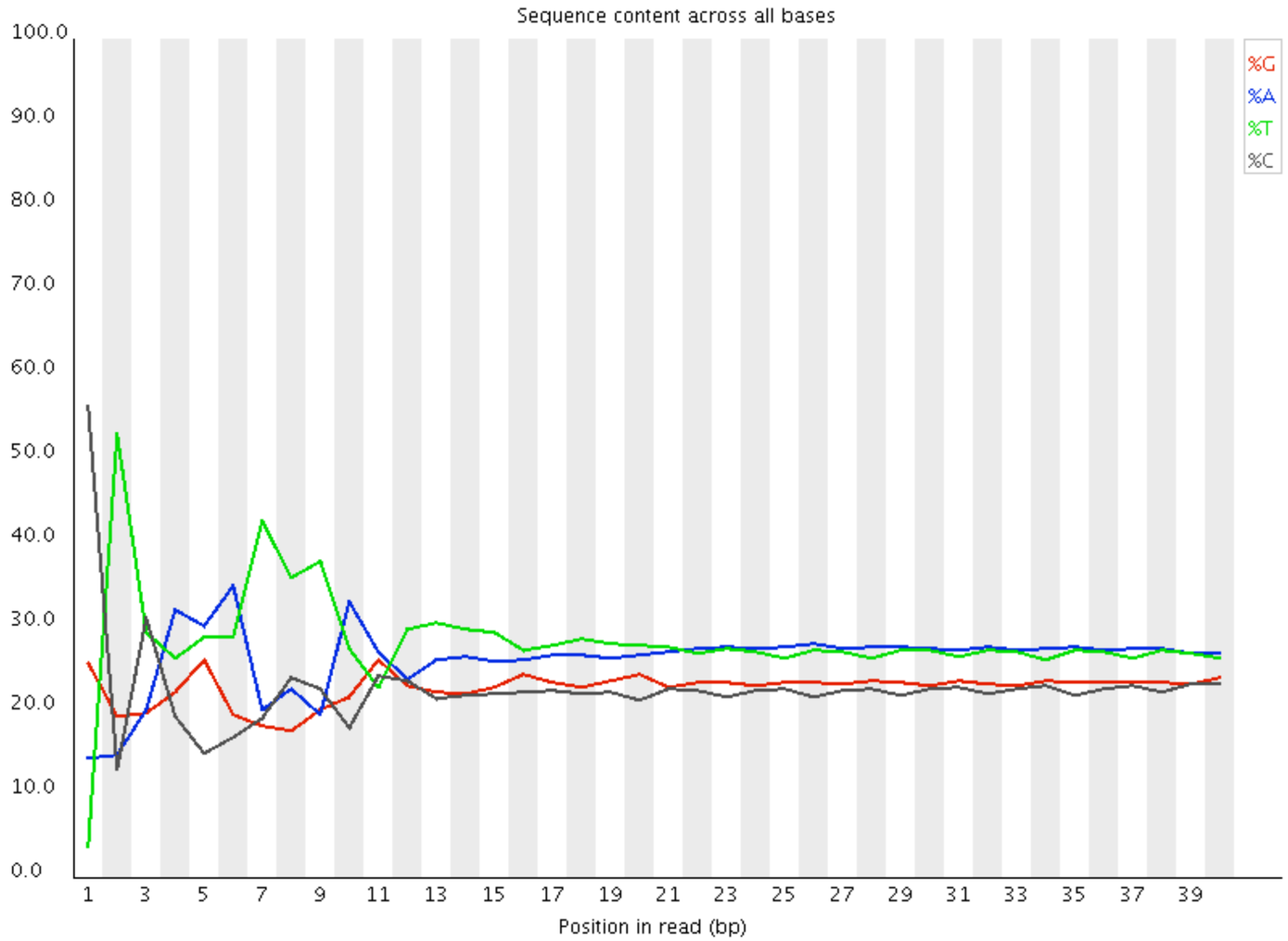


Boundary effects



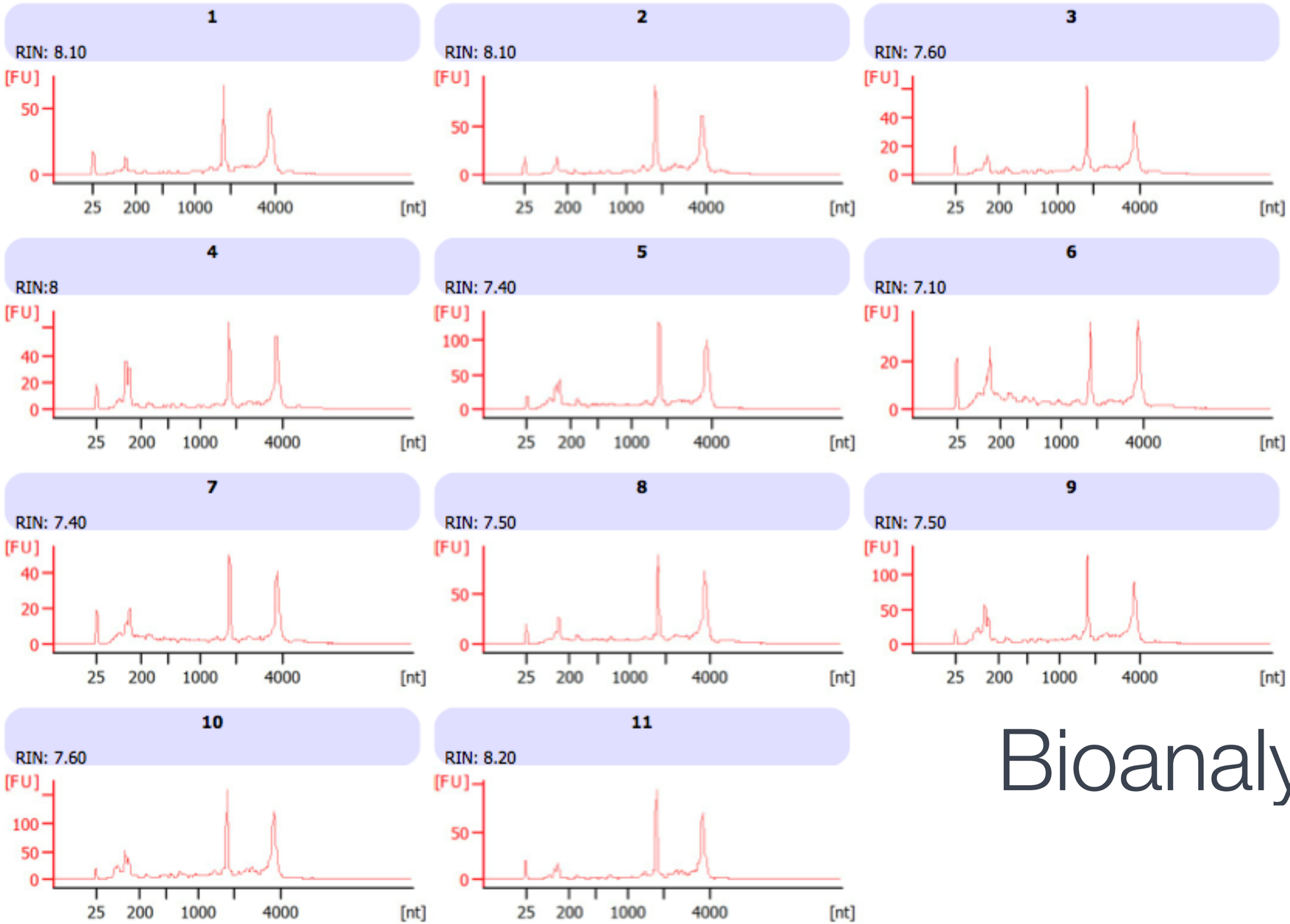
Flurophore accumulation

Illumina: physical/chemical problems



Q&A

See http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010 for more examples



Bioanalyzer

	sequence	count
1	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	482185
151	ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	271724
2	TAATACGACTCACTATAGGGCGAATTGAATTTAGCGGCCGCGAATTCGCC	159936
152	TAATACGACTCACTATAGGGCGAATTGAATTTAGCGGCCGCGAATTCGCC	105273
153	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	46872
3	CTTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC	43212
4	NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	13142

Read Frequency Distribution

QA: filtering

```
> gnl|uv|NGB00105.1:1-219 pCR4-TOPO multiple cloning site
Length=219

 Score =  100 bits (50),  Expect = 9e-19
 Identities = 50/50 (100%), Gaps = 0/50 (0%)
 Strand=Plus/Plus

Query  1   ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC  50
           ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  43  ATTAACCCTCACTAAAGGGACTAGTCCTGCAGGTTTAAACGAATTCGCCC  92
```

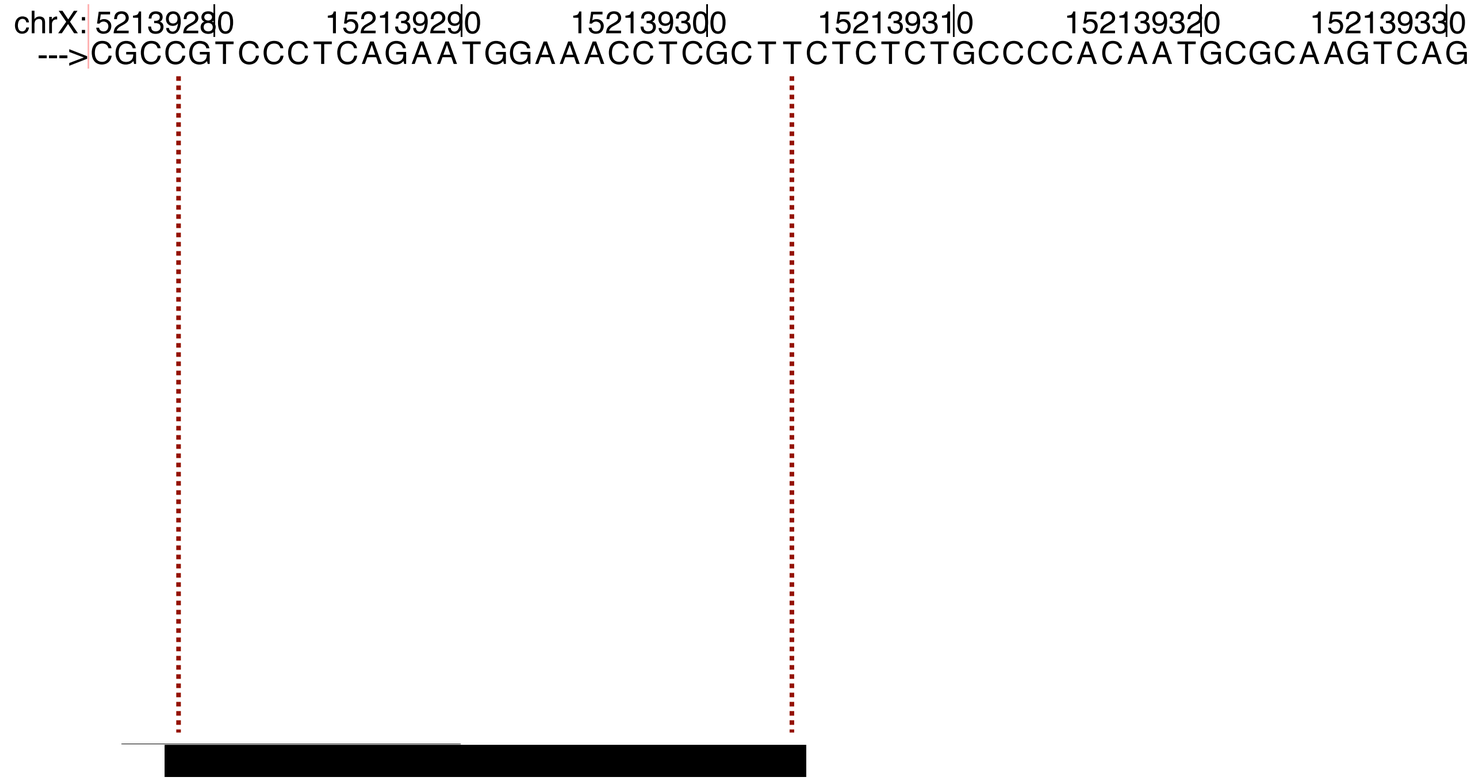
VecBase Screen

QA: filtering

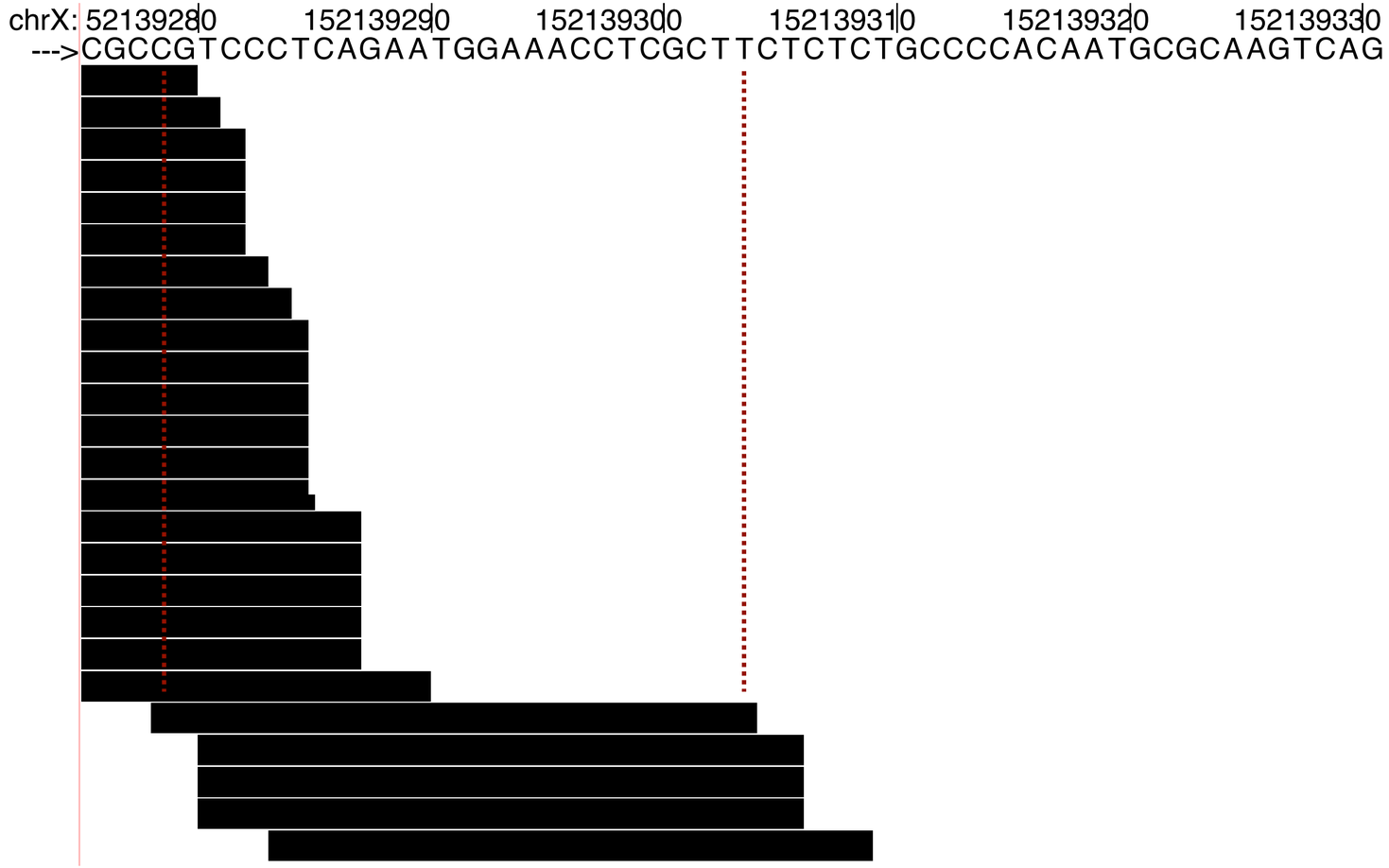
	sequence	count	lane
1051	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	70947	s_5_1_export.txt
451	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	69116	s_4_1_export.txt
601	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	66776	s_6_1_export.txt
301	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	63998	s_3_1_export.txt
751	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	55729	s_7_1_export.txt
151	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	54828	s_2_1_export.txt
901	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	40359	s_8_1_export.txt
1	ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	30880	s_1_1_export.txt
152	ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	30485	s_2_1_export.txt
153	CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	26476	s_2_1_export.txt
2	TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	25600	s_1_1_export.txt
154	GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	25594	s_2_1_export.txt
3	CNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	25063	s_1_1_export.txt
155	TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	24965	s_2_1_export.txt
4	GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	24164	s_1_1_export.txt
302	ANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	22501	s_3_1_export.txt
5	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	20996	s_1_1_export.txt
452	TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	20842	s_4_1_export.txt

QA: filtering

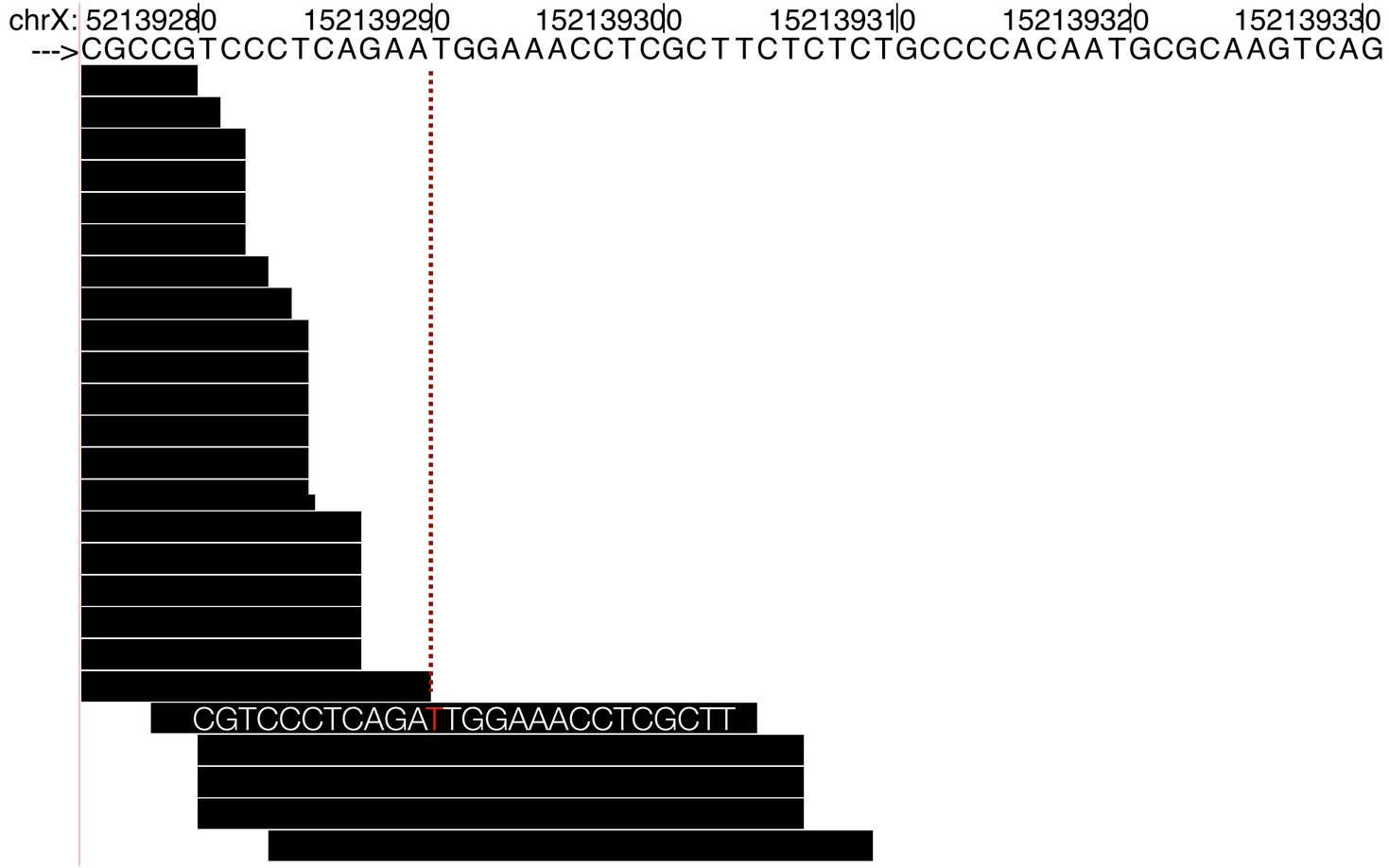
Tool evolution: mapping reads to a genome



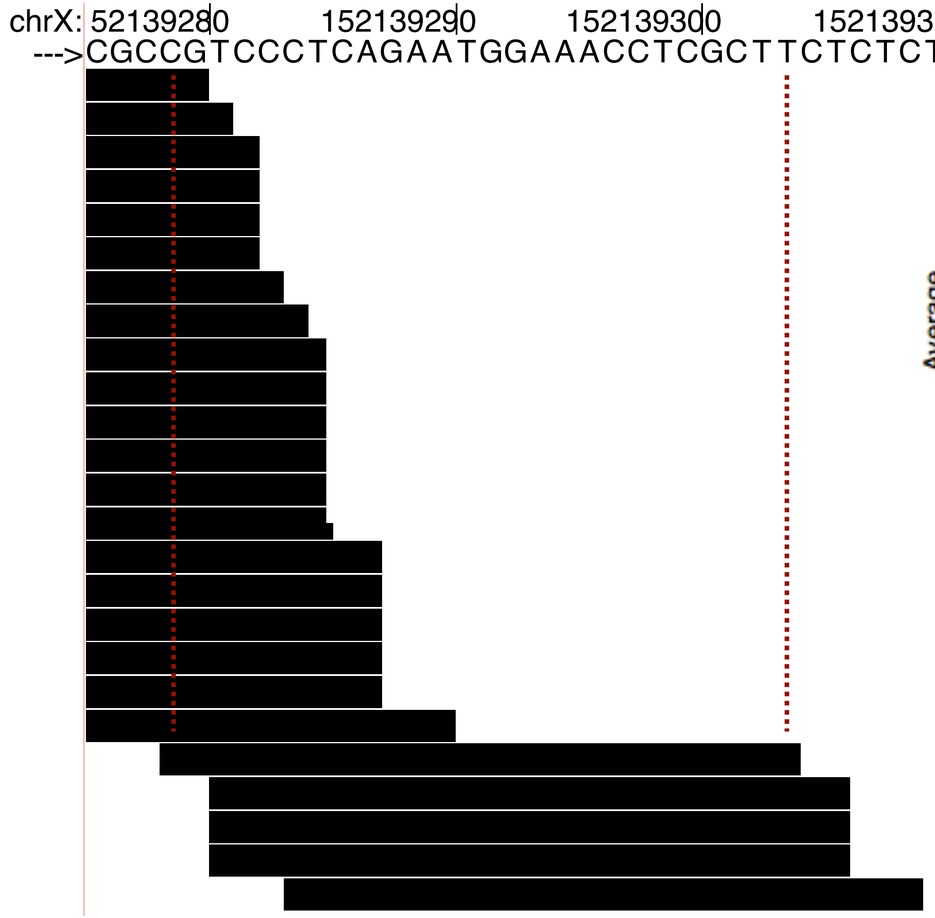
Mapping to a Reference Genome

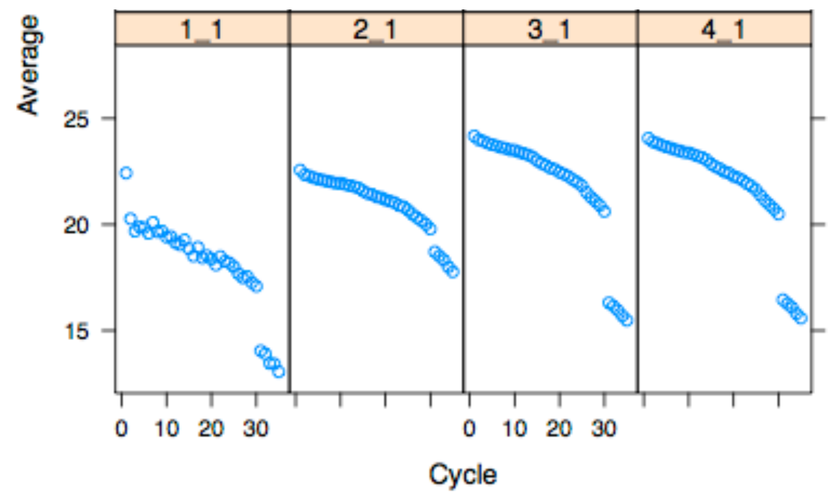


Mapping to a Reference Genome



Mapping to a Reference Genome





Mapping to a Reference Genome

Tool evolution: mapping approaches

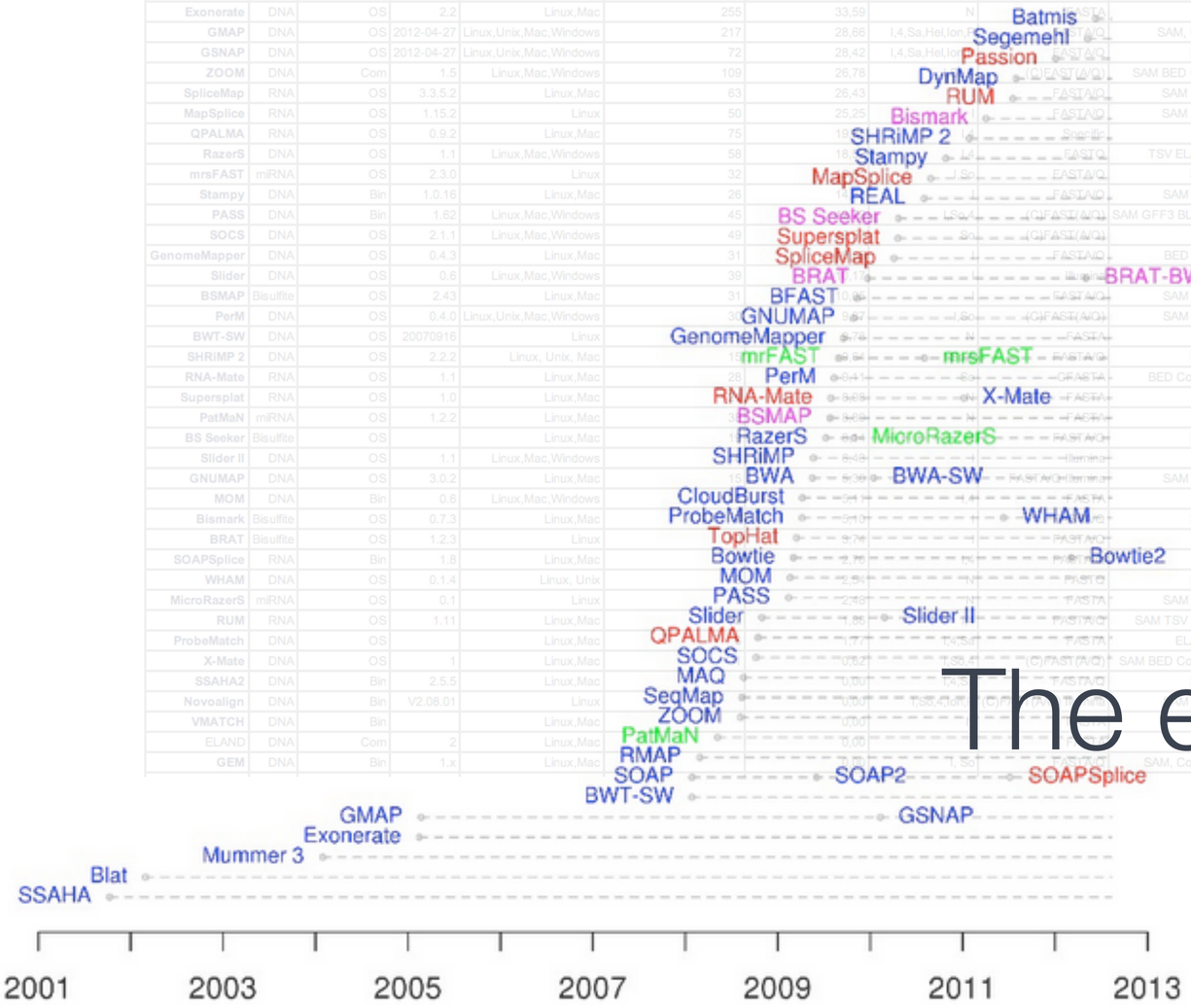
- Variation in algorithm
- Alignment speed
- Memory requirements
- Error tolerance
- ...

Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

The standard tools

Mapper	Data	Availability	Version	O.S.	Number Citations	Citations/Years	Seq.Plat.	Input	Output	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align. Reported	Alignment	Parallel	QA	PE	Splicing
Bowtie	DNA	OS	0.12.7	Linux,Mac,Windows	1168	335,04	I,So,4,Sa,P	(C)FAST(A/Q)	SAM TSV	4	1K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N
Blat	DNA	OS	34	Linux,Mac	2844	268,37	N	FASTA	TSV BLAST	11	5000K	Score	Score	Y	B	L	N	N	N	De novo
MAQ	DNA	OS	0.7.1	Linux,Mac	957	237,27	I,So	(C)FAST(A/Q)	TSV	8	63	Y	Y	N			N	Y	Y	N
BWA	DNA	OS	0.6.2	Linux,Mac,Windows	738	225,15	I,So,4,Sa,P	FASTA/Q	SAM	4	200	Y	8	Y	R,S	G	SM	Y	Y	N
TopHat	RNA	OS	1.4.1	Linux,Mac	389	112,66	I	FASTA/Q, GFF	BAM	-	-	2	0	N	B,S	-	SM	Y	Y	De novo
SOAP	DNA	OS	1.11	Linux,Mac	451	98,04	I	FASTA/Q	TSV	7	60	5	3	N	B,R S		SM	N	Y	N
SOAP2	DNA	OS	2.21	Linux	294	90,93	I	FASTA/Q	SAM TSV	27	1K	2	0	Y	A,B,R	L	SM	N	Y	N
Mummer 3	DNA	OS	3.23	Linux,Mac	683	78,93	N	FASTA	TSV	10	*	Y	Y	Y	A,B	G	N	N	N	N
BWA-SW	DNA	OS	0.6.2	Linux,Mac,Windows	160	61,41	I,So,4,Sa,P	FASTA/Q	SAM	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N
mrFAST	miRNA	OS	2.1.0.4	Linux	158	52,86	I	FASTA/Q	SAM	25	300	Score	6	N	A,B	G	N	N	Y	N
SHRiMP	DNA	OS	1.3.2	Linux,Mac	155	47,45	I,So,4,Hel	(C)FAST(A/Q)	TSV	14	1K	Score	Score	Y	B,S	G	SM	N	Y	N
SSAHA	DNA	OS	3.1	Linux,Mac	483	43,94	N	FASTA/Q	TSV	15	*	Y	Y	Y	B,S	G L	N	N	N	N
CloudBurst	DNA	OS	1.1	Linux,Mac,Windows	146	43,08	N	FASTA	TSV		1K	Y	Y	Y	A,B	G	Cloud	N	N	N
RMAP	DNA	OS	2.05	Linux,Mac	162	35,89	I,So,4	(C)FAST(A/Q)	BED	11	10K	Y	0	N	B,S		N	Y	Y	N
SeqMap	DNA	OS	1.013	Linux,Mac	142	35,04	I	FASTA	ELAND	15	500	5	3	N	A		SM	N	N	N
BFAST	DNA	OS	0.7.0	Linux,Mac	94	33,74	I,So,4, Hel	(C)FAST(A/Q)	SAM TSV		*	Y	Y	Y	B,R,U	G	SM	N	Y	N
Exonerate	DNA	OS	2.2	Linux,Mac	255	33,59	N	FASTA	TSV	20	*	Score	Score	Y	B,S	G L	N	N	N	De novo
GMAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	217	28,66	I,4,Sa,Hel,Ion	FASTA/Q	SAM, GFF	8	*	Y	Y	Y	B	G L	SM	N	N	De novo
GSNAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	72	28,42	I,4,Sa,Hel,Ion	FASTA/Q	SAM	8	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and de novo
ZOOM	DNA	Com	1.5	Linux,Mac,Windows	109	26,78	I	(C)FAST(A/Q)	SAM BED GFF	12	240	Y	Y	N	B,U,S	G	SM/DM	Y	Y	N
SpliceMap	RNA	OS	3.3.5.2	Linux,Mac	63	26,43	I	FASTA/Q	SAM BED	-	-	0.1		Y	A	-	SM	N	Y	Lib and/or de novo
MapSplice	RNA	OS	1.15.2	Linux	50	25,25	I	FASTA/Q	SAM BED	-	-	3		Y	B	-	SM	N	Y	De novo
QPALMA	RNA	OS	0.9.2	Linux,Mac	75	19,23	I	Specific	TSV	-	-	Y	Y	Y	B	L	N	Y	N	Lib and de novo
RazerS	DNA	OS	1.1	Linux,Mac,Windows	58	18,57	I	FASTA/Q	TSV ELAND	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N
mrsFAST	miRNA	OS	2.3.0	Linux	32	16,00	I,So	FASTA/Q	SAM	25	200	Y	0	N	A	G	N	N	Y	N
Stampy	DNA	Bin	1.0.16	Linux,Mac	26	13,00	I	FASTA/Q	SAM TSV	4	4K	0.15	30	N	B,R,S	G	N	Y	Y	N
PASS	DNA	Bin	1.62	Linux,Mac,Windows	45	11,25	I,So,4	(C)FAST(A/Q)	SAM GFF3 BLAST	23	1K	Y	Y	Y	A,B	G	SM	Y	Y	De novo
SOCS	DNA	OS	2.1.1	Linux,Mac,Windows	49	10,89	So	(C)FAST(A/Q)	TSV		64	Y	0	N	A,B		SM	Y	N	N
GenomeMapper	DNA	OS	0.4.3	Linux,Mac	31	10,33	I	FASTA/Q	BED TSV	12	2K	10	10	Y	A,B,R	G	SM	N	N	N
Slider	DNA	OS	0.6	Linux,Mac,Windows	39	10,00	I	(C)FAST(A/Q)	SAM		62	3	0	N	B,S		N	Y	Y	N
BSMAP	Bisulfite	OS	2.43	Linux,Mac	31	10,33	I	FASTA/Q	SAM TSV	8	144	15	0	N	B,U,S		SM	N	Y	N
PerM	DNA	OS	0.4.0	Linux,Unix,Mac,Windows	30	9,67	I,So	(C)FAST(A/Q)	SAM TSV	20	128	9	0	Y	A,U	G	DM	Y	Y	N
BWT-SW	DNA	OS	20070916	Linux	9,76		N	FASTA	TSV		1K	Score	Score	Y	A		N	N	N	N
SHRiMP 2	DNA	OS	2.2.2	Linux, Unix, Mac	19	9,50	I	FASTA/Q	SAM	30	1K	Y	Score	N	B,U,S	G	SM	Y	Y	N
RNA-Mate	RNA	OS	1.1	Linux,Mac	28	9,33	I	CSFASTA	BED Counts	-	-	Y	0	N	S	-	DM	Y	N	Lib
Supersplat	RNA	OS	1.0	Linux,Mac	9,00		I,So	(C)FAST(A/Q)	TSV			0	0	Y	A,U	G	N	N	N	De novo
PatMaN	miRNA	OS	1.2.2	Linux,Mac	30	8,57	I	FASTA/Q	TSV	1	*	Y	Y	N	A	G	N	N	N	N
BS Seeker	Bisulfite	OS		Linux,Mac	8,00		I	FASTA/Q	SAM	-	-	3	0	N	U	-	SM	Y	N	N
Slider II	DNA	OS	1.1	Linux,Mac,Windows	8,00		I	FASTA/Q	TSV		93	Y		N	B,S		N	N	Y	N
GNUMAP	DNA	OS	3.0.2	Linux,Mac	17	8,50	I,So	(C)FAST(A/Q)	SAM TSV	16	1K	Score	Score	Y	B	G	SM/DM	Y	N	N
MOM	DNA	Bin	0.6	Linux,Mac,Windows	8,00		I	FASTA/Q	TSV			Y	0	N	A	L	SM	N	Y	N
Bismark	Bisulfite	OS	0.7.3	Linux,Mac	16	8,00	I	FASTA/Q	SAM	16	10K	Score	Score	N	U	-	SM	Y	Y	N
BRAT	Bisulfite	OS	1.2.3	Linux	7,67		I	FASTA/Q	TSV			Y	0	N			N	N	Y	N
SOAPSplice	RNA	Bin	1.8	Linux,Mac	7,67		I	FASTA/Q	TSV	13	3K	5	2	Y	U	-	SM	Y	Y	De novo
WHAM	DNA	OS	0.1.4	Linux, Unix	7,50		I	FASTA/Q	SAM	5	128	5	3	N	A,B,R,U,S	G	N	Y	Y	De novo
MicroRazerS	miRNA	OS	0.1	Linux	7,50		I	FASTA/Q	SAM TSV	10	*	Score	0	N	S	G	N	N	N	N
RUM	RNA	OS	1.11	Linux,Mac	7,33		I	FASTA/Q	SAM TSV BED	-	-	Y	Y	Y	B	-	SM	N	Y	De novo
ProbeMatch	DNA	OS		Linux,Mac	7,00		I	FASTA/Q	ELAND	36	50	3	Y	N	A,B		N	N	N	N
X-Mate	DNA	OS	1	Linux,Mac	7,00		I	(C)FAST(A/Q)	SAM BED Counts	-	-	Y	0	N	S		DM	Y	N	Lib
SSAHA2	DNA	Bin	2.5.5	Linux,Mac	6,67		I,So,4,Hel	FASTA/Q	SAM	15	48K	Score	Score	N	B,S	L	N	N	Y	N
Novoalign	DNA	Bin	V2.08.01	Linux	6,67		I,So,4,Hel	(C)FAST(A/Q)	SAM	30	300	8	Score	N	A,B,R,U,S	G	SM/DM/Cloud	Y	Y	Lib
VMATCH	DNA	Bin		Linux,Mac	6,67		I	FASTA/Q	TSV			Y	Score	Y	A,B,S	G L	N	N	N	N
ELAND	DNA	Com	2	Linux,Mac	6,67		I	FASTA/Q	TSV	32		2	0	N	B		N	N	N	N
GEM	DNA	Bin	1.x	Linux,Mac	6,67		I	FASTA/Q	SAM, Counts		*	Y	Y	Y	A,S	G	SM	Y	Y	Lib and de novo



The extended selection

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

SAM/BAM

Sequence/**A**lignment **F**ormat

Read name

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239
```

```
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBBDBDB=?FFECFACCFFHHH>09C
```

Read sequence

Read quality

SAM/BAM

Courtesy of Nick Croucher, HSPH

Bitwise flag

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

SAM/BAM

Courtesy of Nick Croucher, HSPH

Mapping position Mapping quality

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

SAM/BAM

Courtesy of Nick Croucher, HSPH

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239                     
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

Alignment

SAM/BAM

Courtesy of Nick Croucher, HSPH

M: match
I: insertion relative to reference
D: deletion relative to reference

```
                                  1 1 1 1 1 1 1 1 1 1
RefPos: 1 2 3 4 5 6 7   8 9 0 1 2 3 4 5 6 7 8 9
----------------------------------------------------
Ref:    C C A T A C T   G A A C T G A C T A A C
Read:         A C T A G A A   T G G C T
```

CIGAR strings

Courtesy of Nick Croucher, HSPH

M: match
I: insertion relative to reference
D: deletion relative to reference

```
                                  1 1 1 1 1 1 1 1 1 1
RefPos: 1 2 3 4 5 6 7   8 9 0 1 2 3 4 5 6 7 8 9
----------------------------------------------------
Ref:    C C A T A C T   G A A C T G A C T A A C
Read:         A C T A G A A   T G G C T
```

CIGAR string: 3M1I3M1D5M

CIGAR strings

Courtesy of Nick Croucher, HSPH

Distance to mate pair

```
M00628:11:000000000-A1P5L:1:1112:26953:13136
163     CP000921        20      60      149M
=       108     239
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT
?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

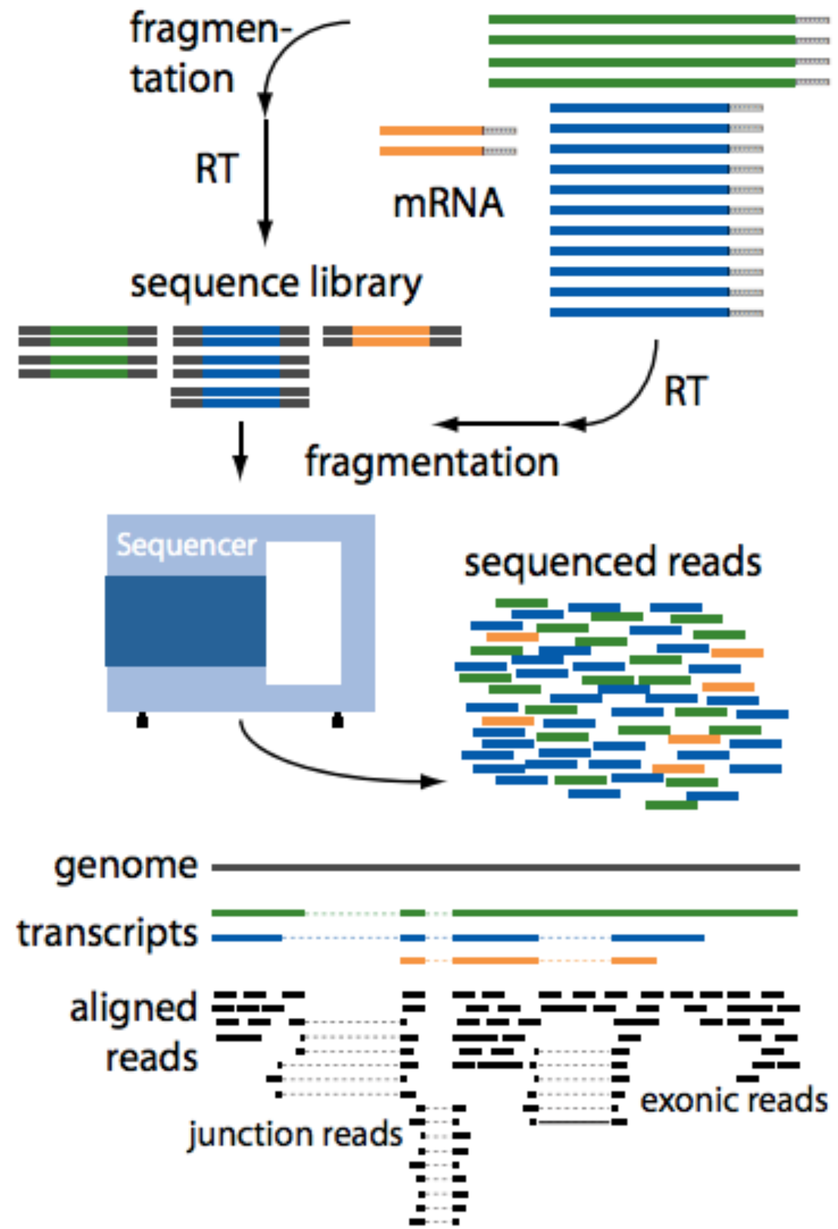
Mate mapped to same reference?

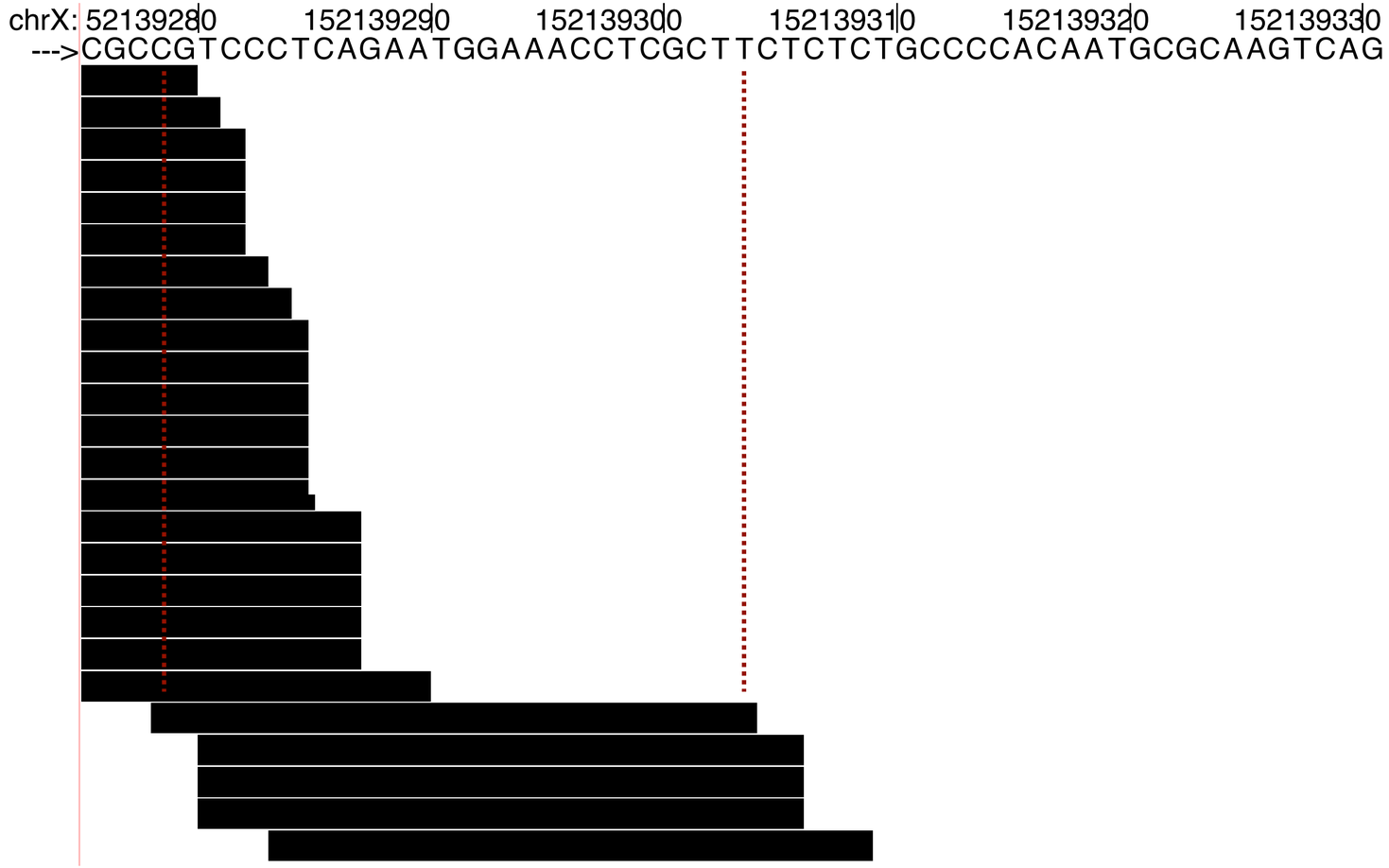
Total length

SAM/BAM

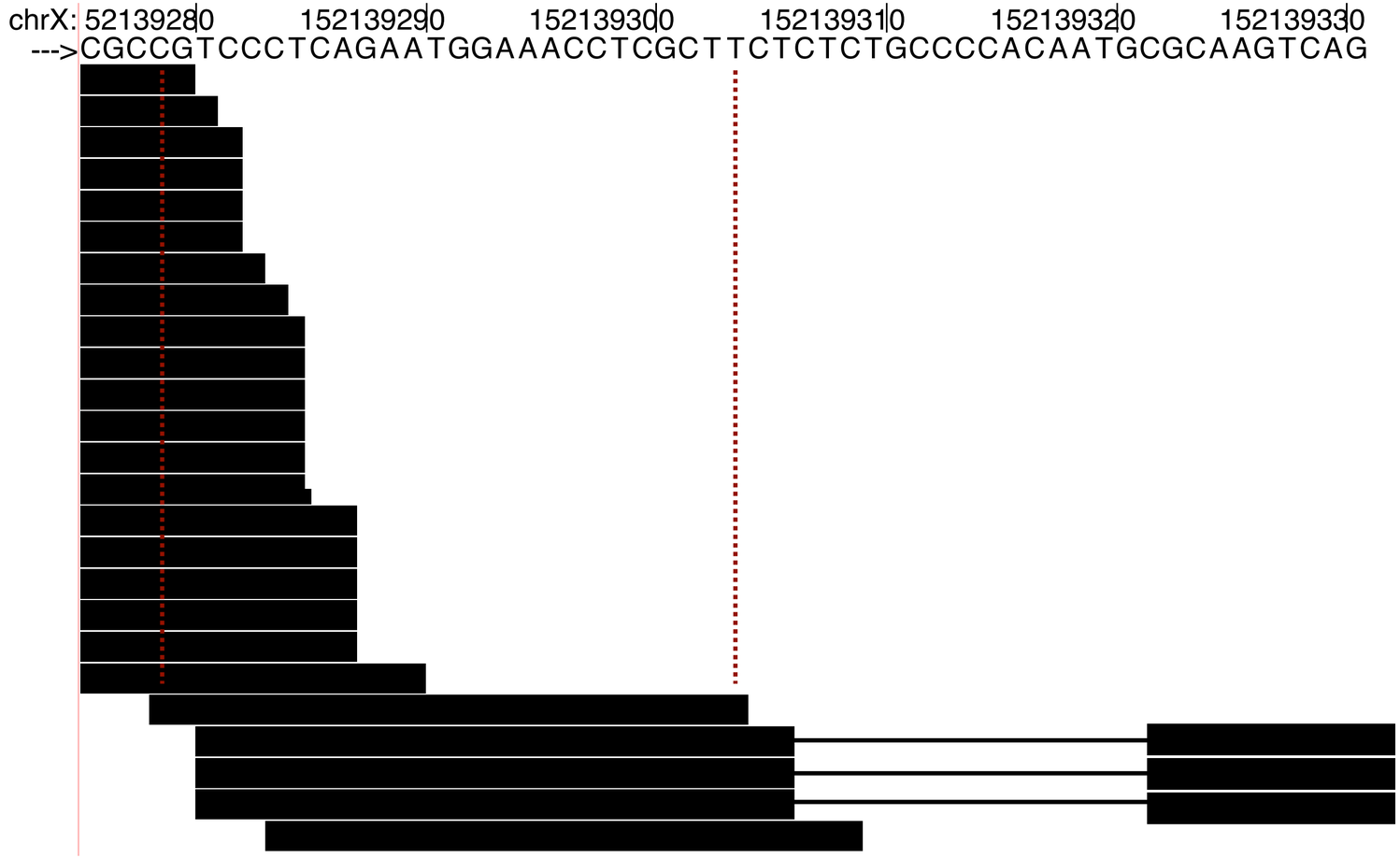
Courtesy of Nick Croucher, HSPH

RNA-seq

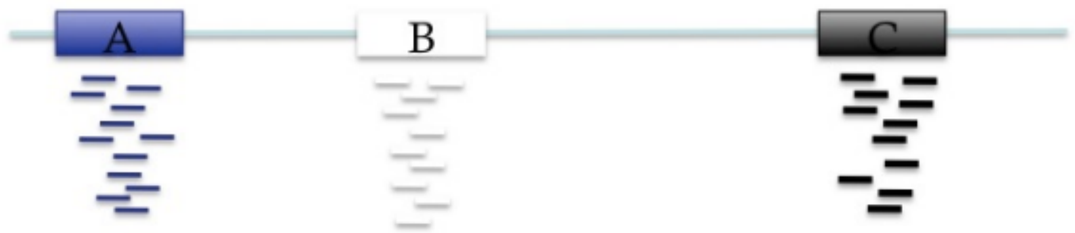

fragmen-
tation
RT
mRNA
sequence library
RT
fragmentation
Sequencer
sequenced reads
genome
transcripts
aligned
reads
exonic reads
junction reads



Mapping to a Reference Genome

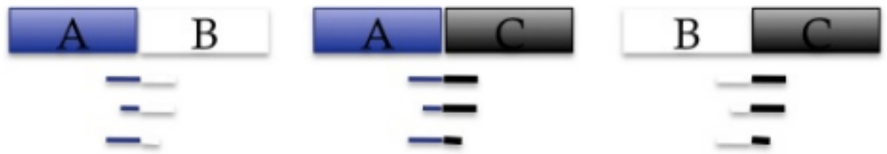


Mapping to a Reference Genome



identify candidate exons via genomic mapping

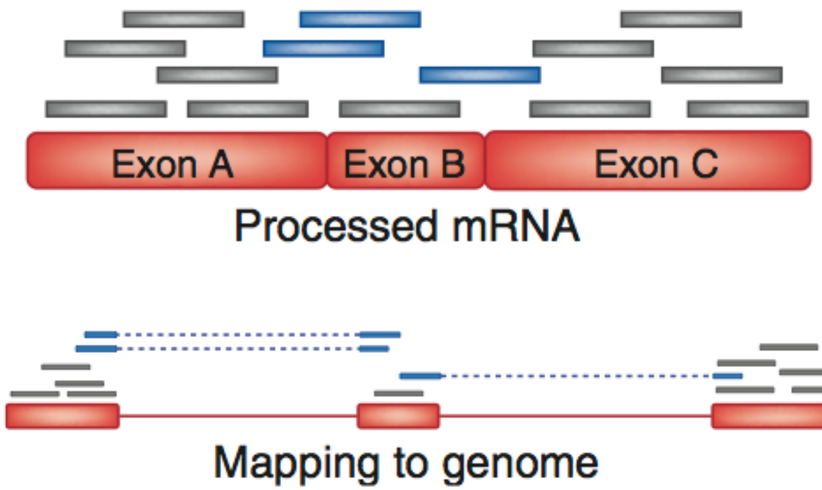
Generate possible pairings of exons

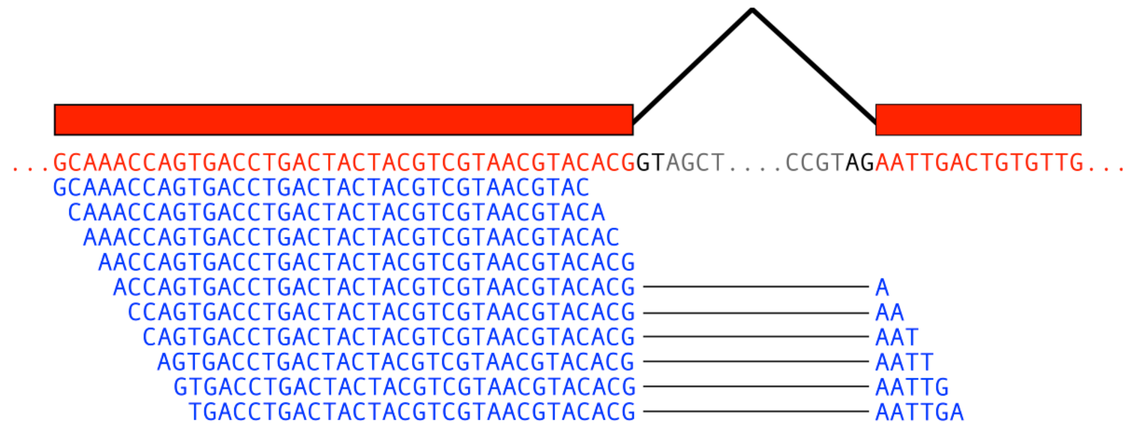


Align reads to possible junctions

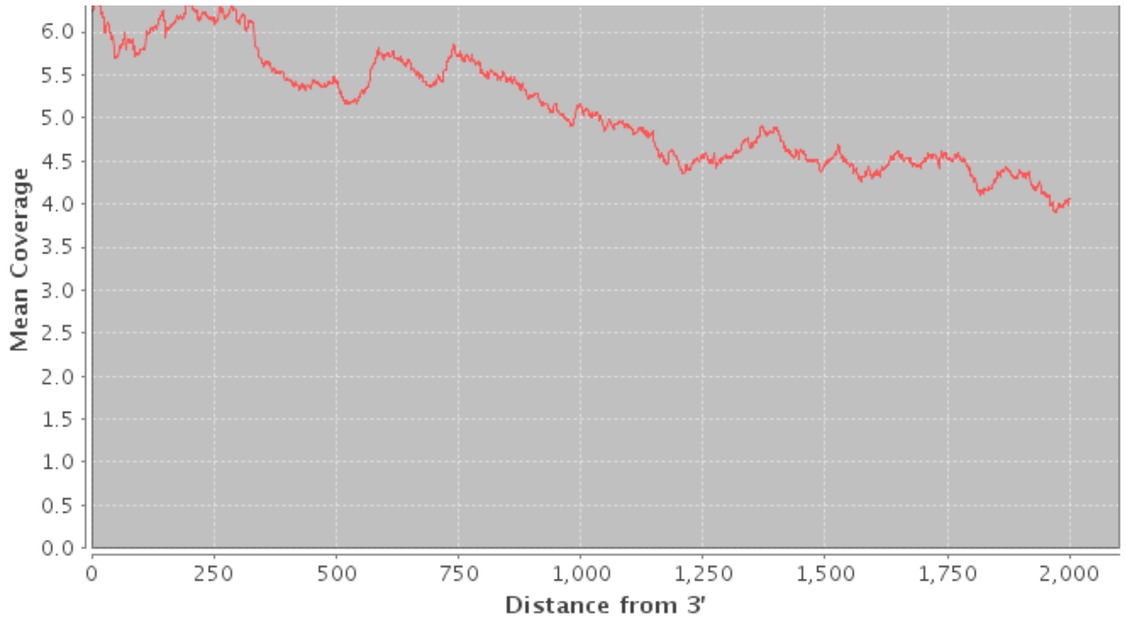
(Trapnell, 2010)

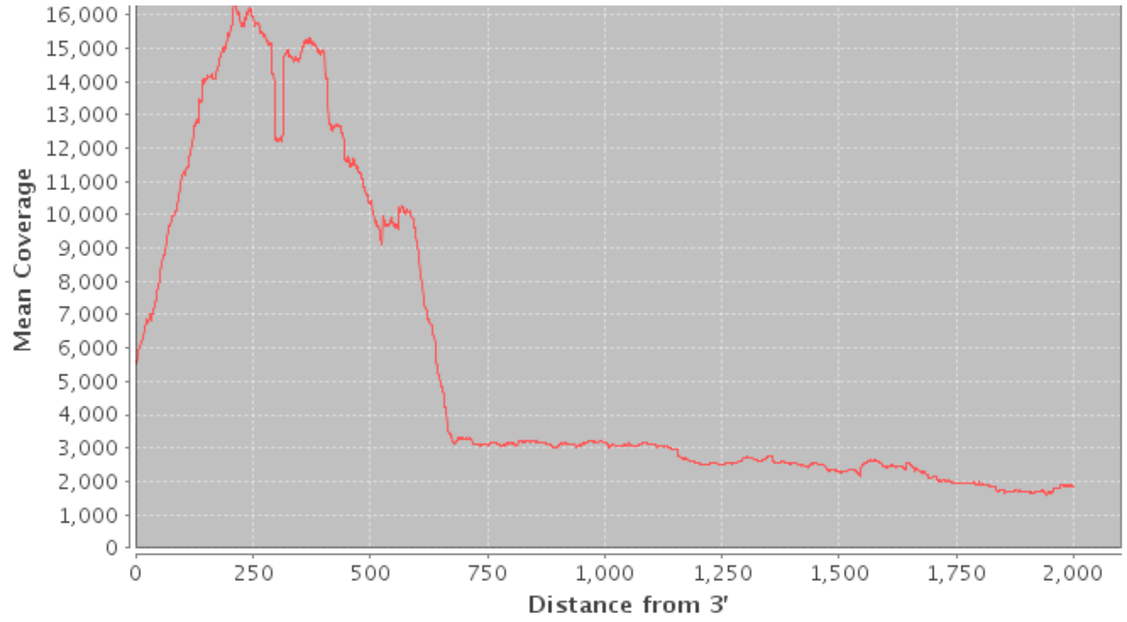
TopHat alignment concept





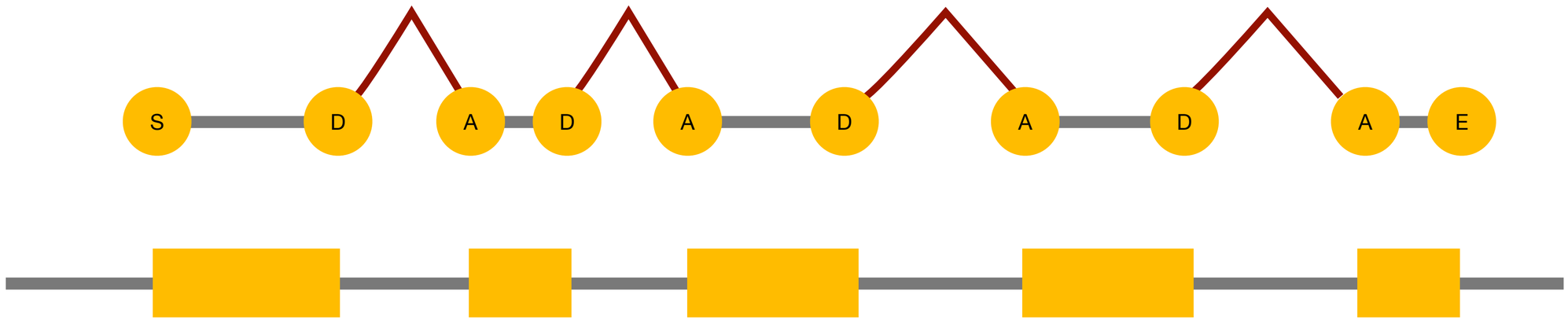
Transcript discovery



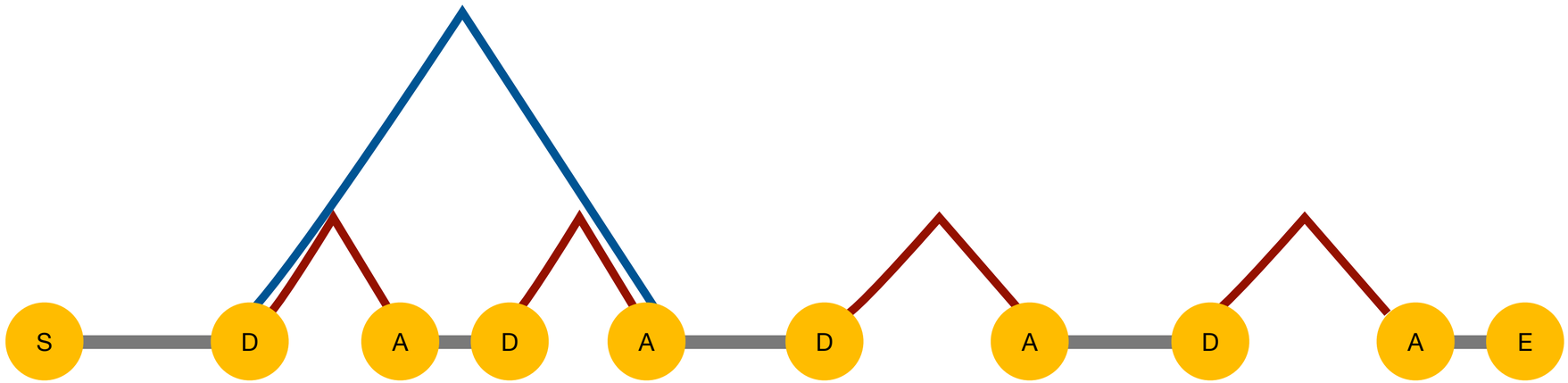


Additional QC

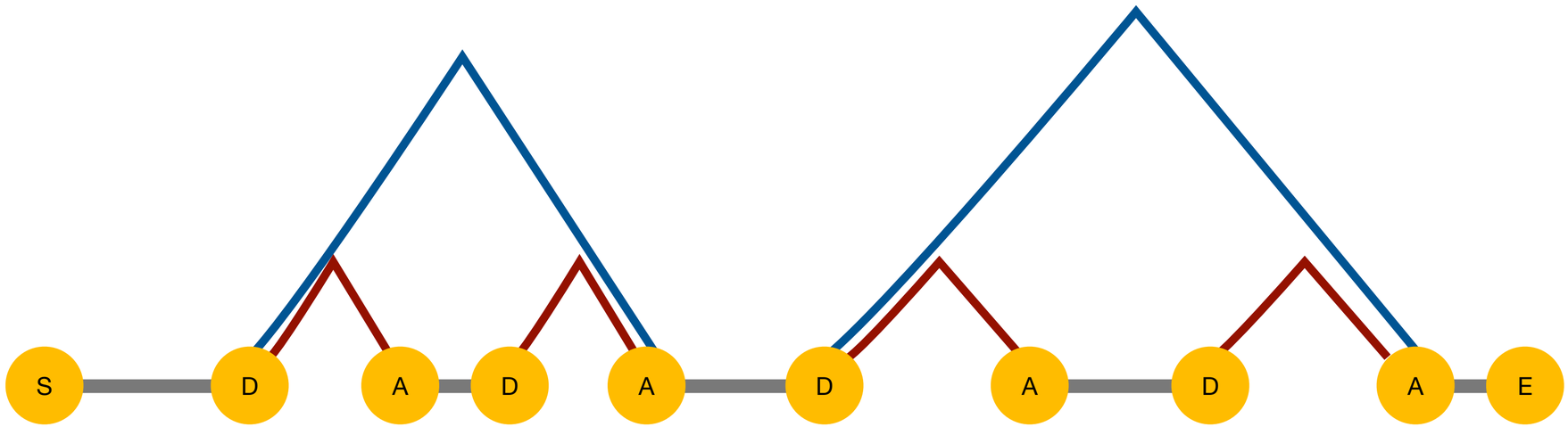
Splice graphs



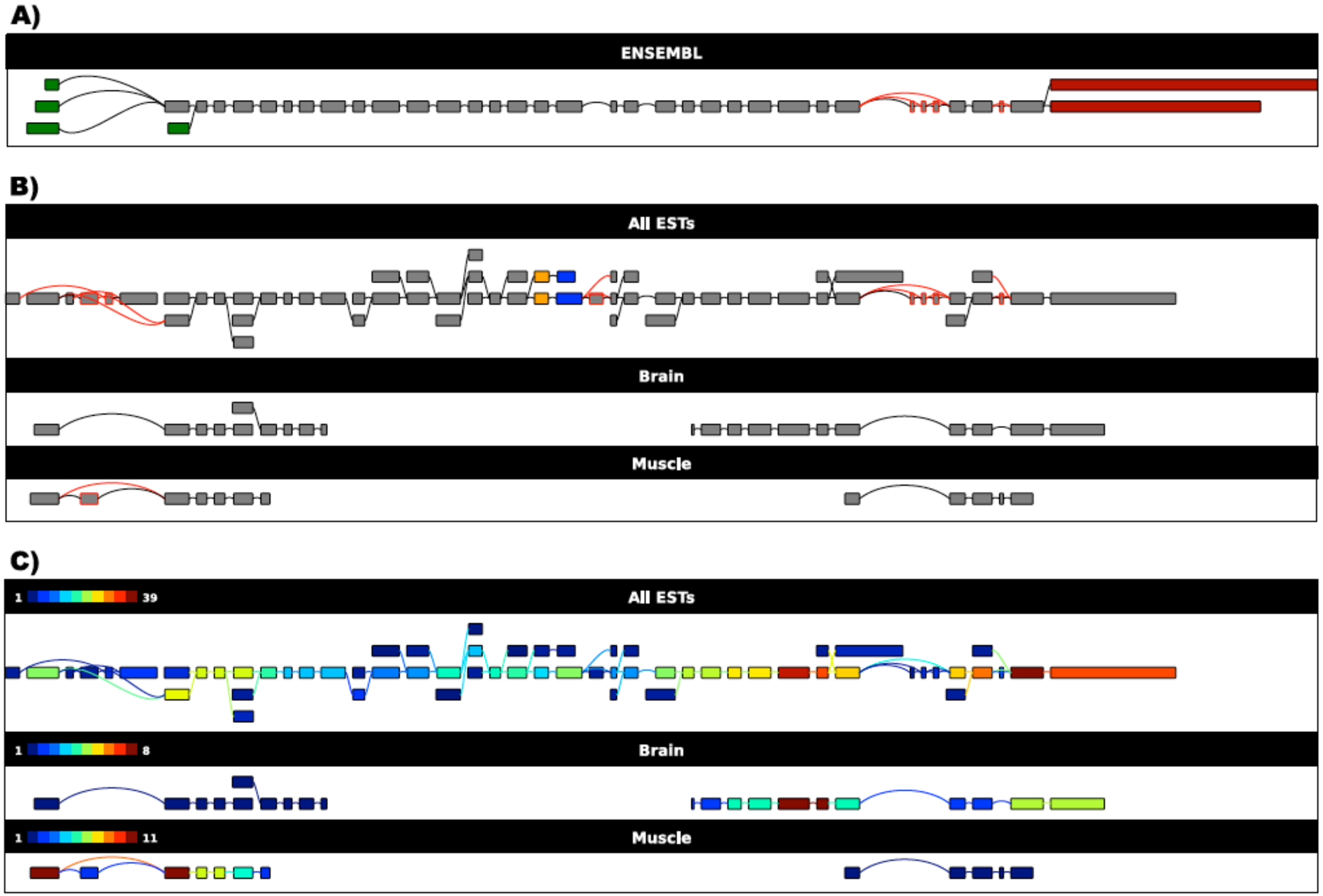
Splice graphs



Splice graphs

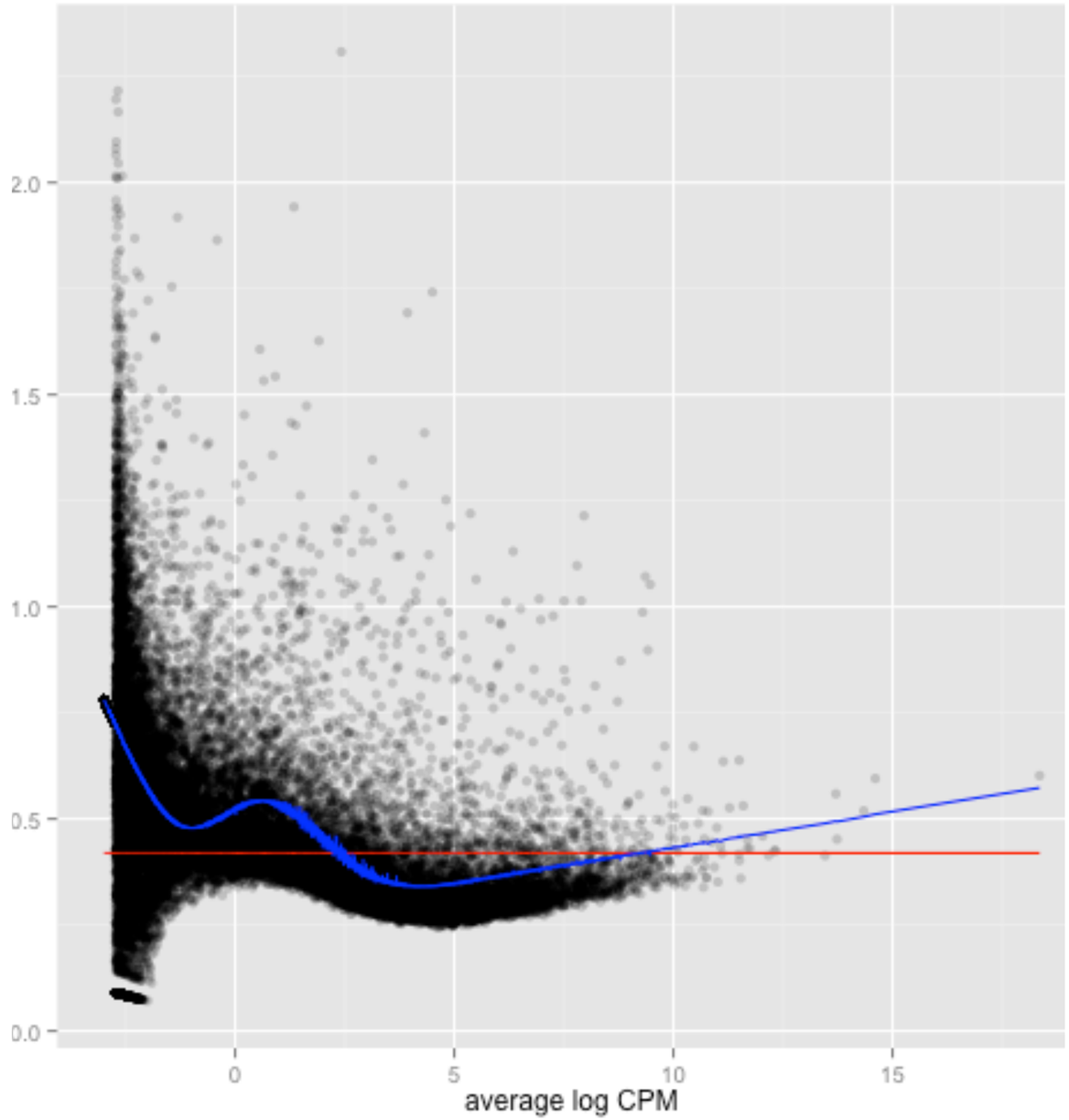


Splice graphs


A)
ENSEMBL
B)
All ESTs
Brain
Muscle
C)
1
39
All ESTs
1
8
Brain
1
11
Muscle

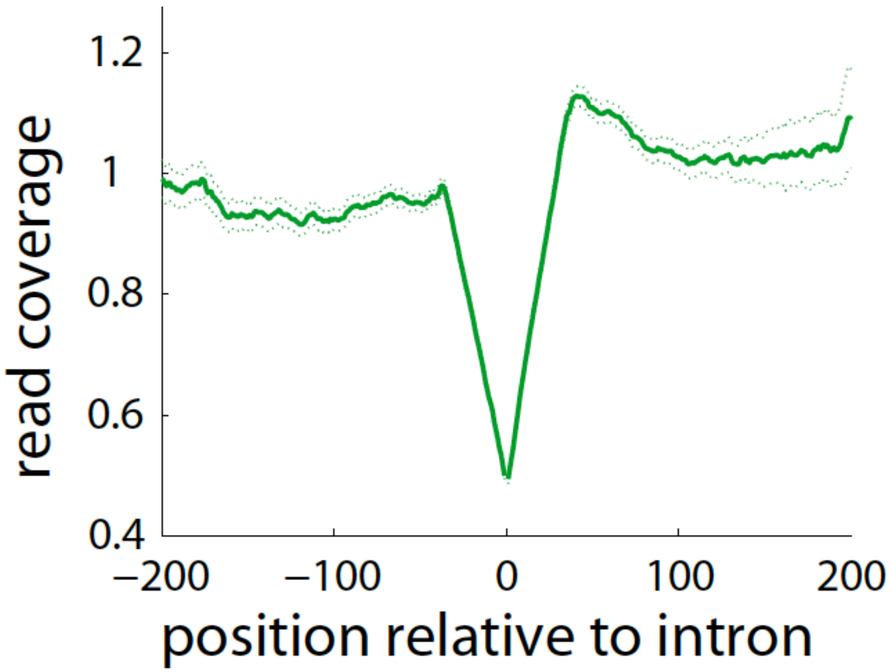
Quantification

Count normalization



Biases

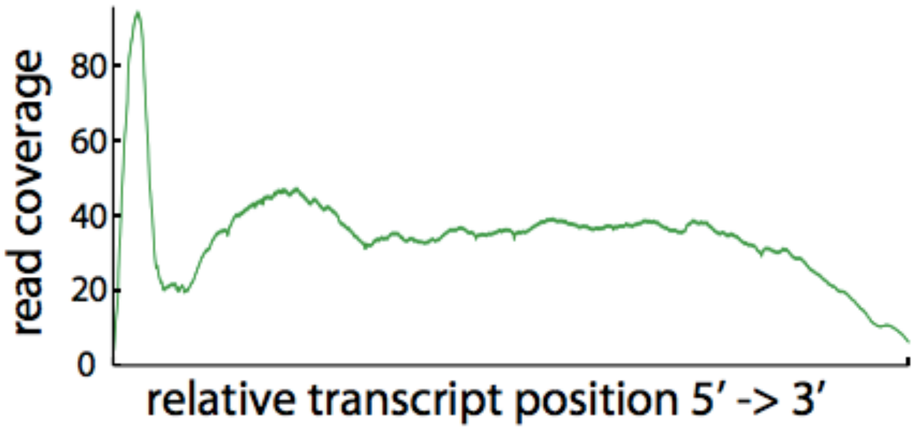
Read mapping

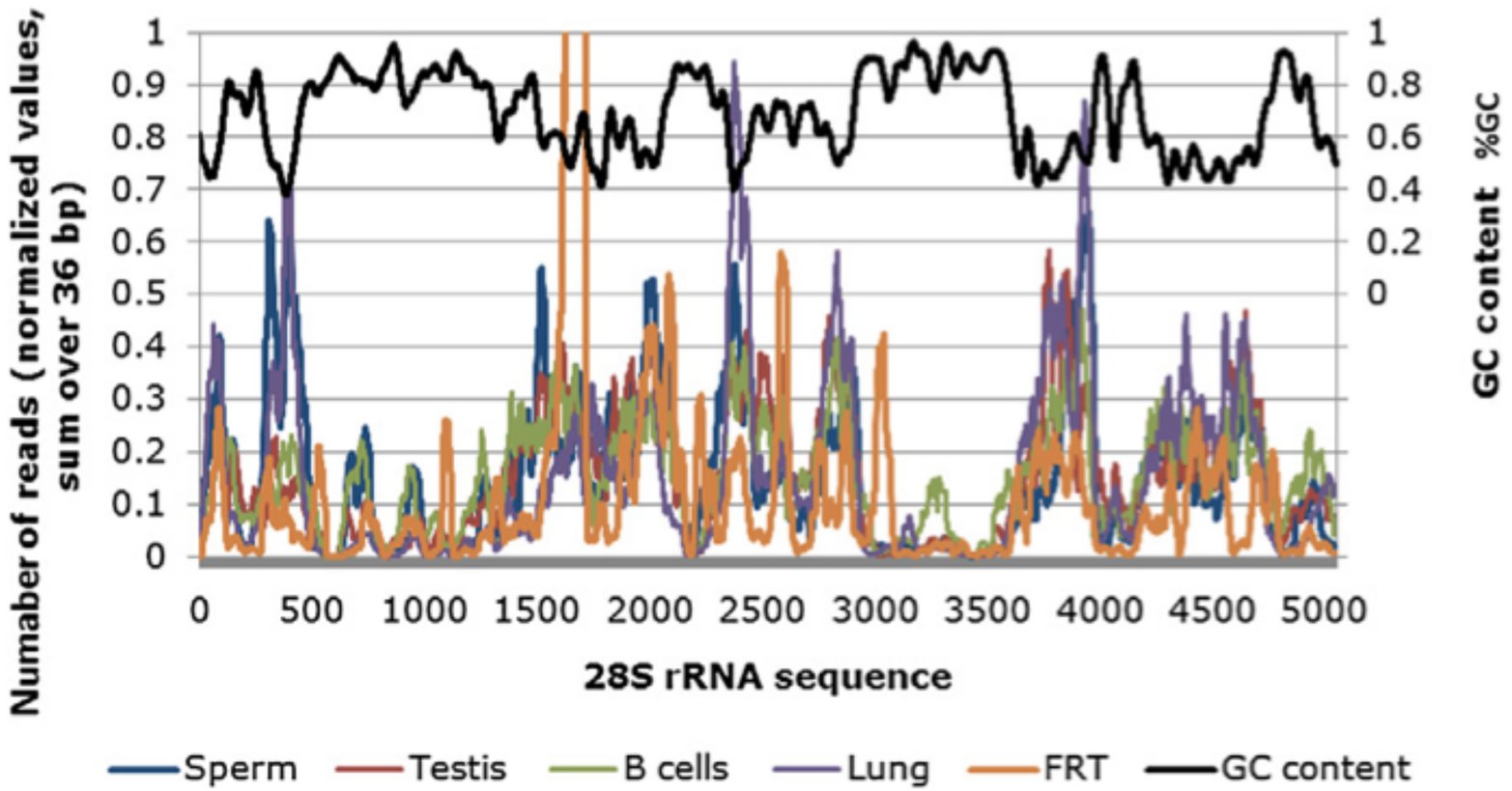


Biases

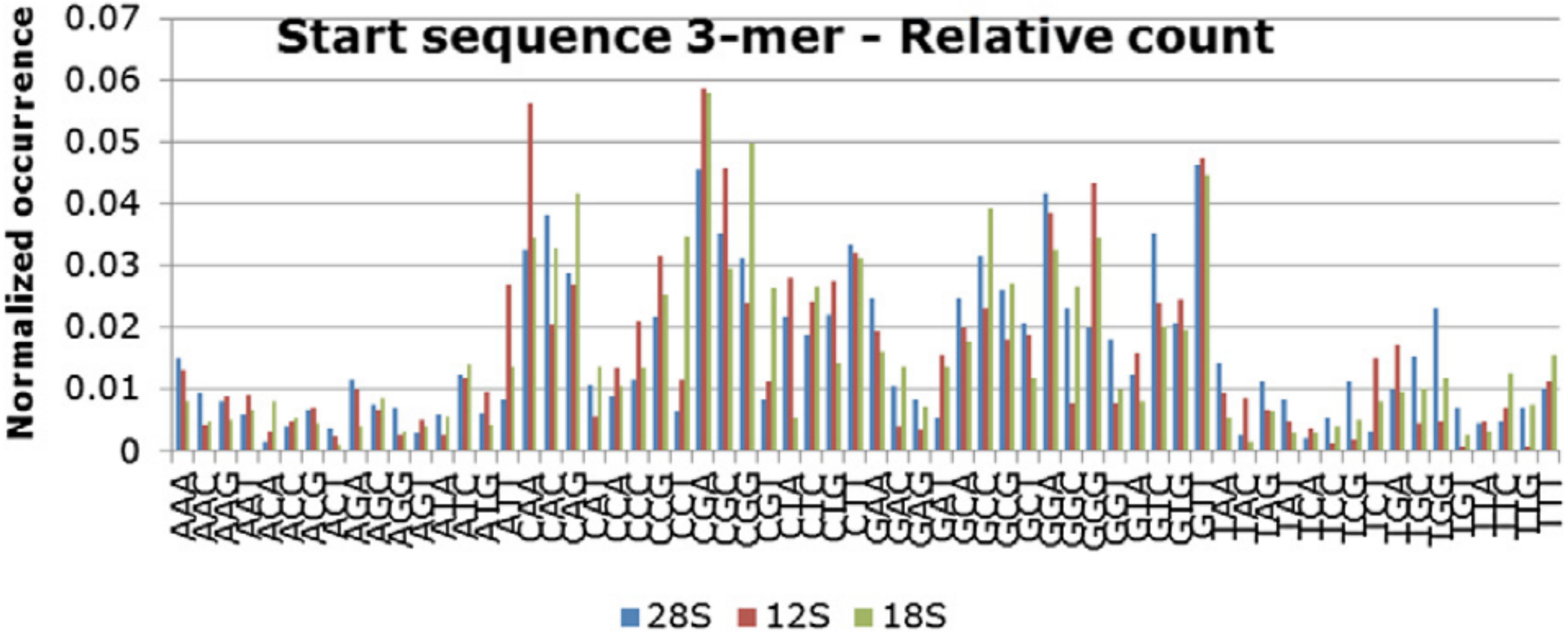
Library construction

Transcript length

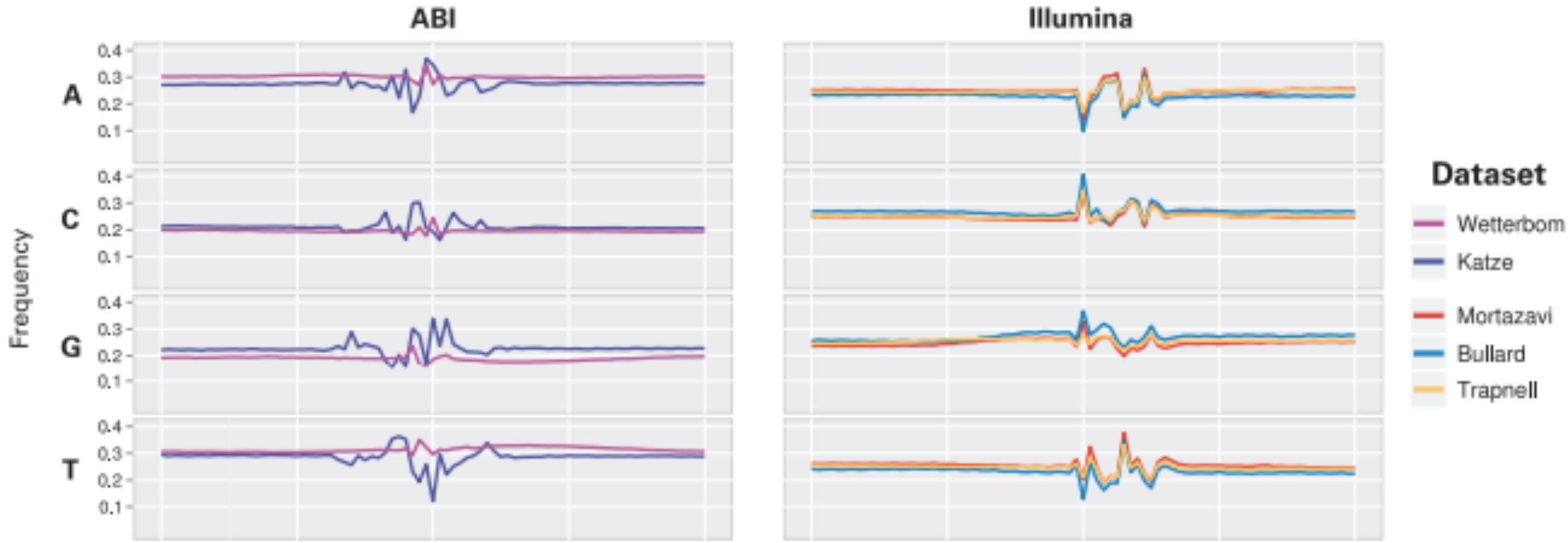




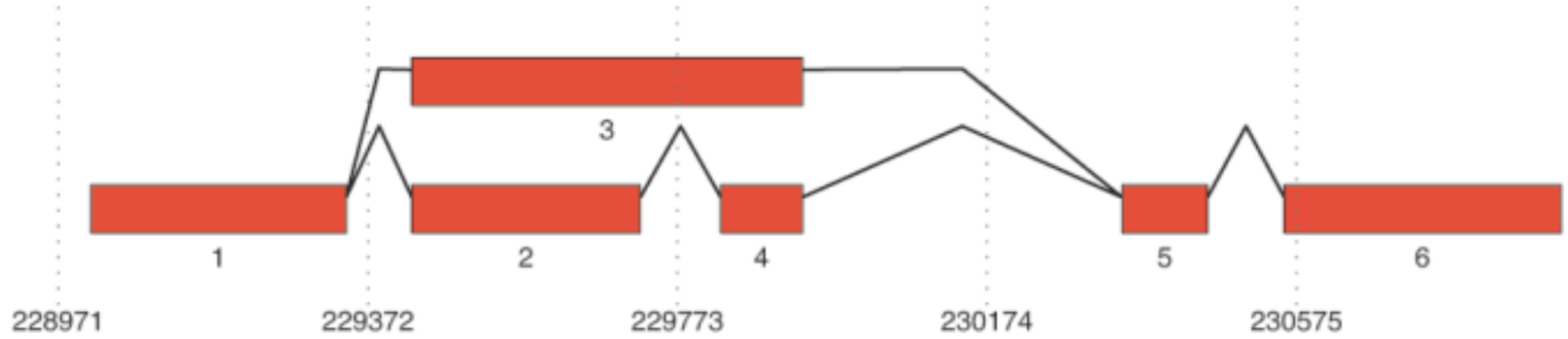
Nucleotide composition

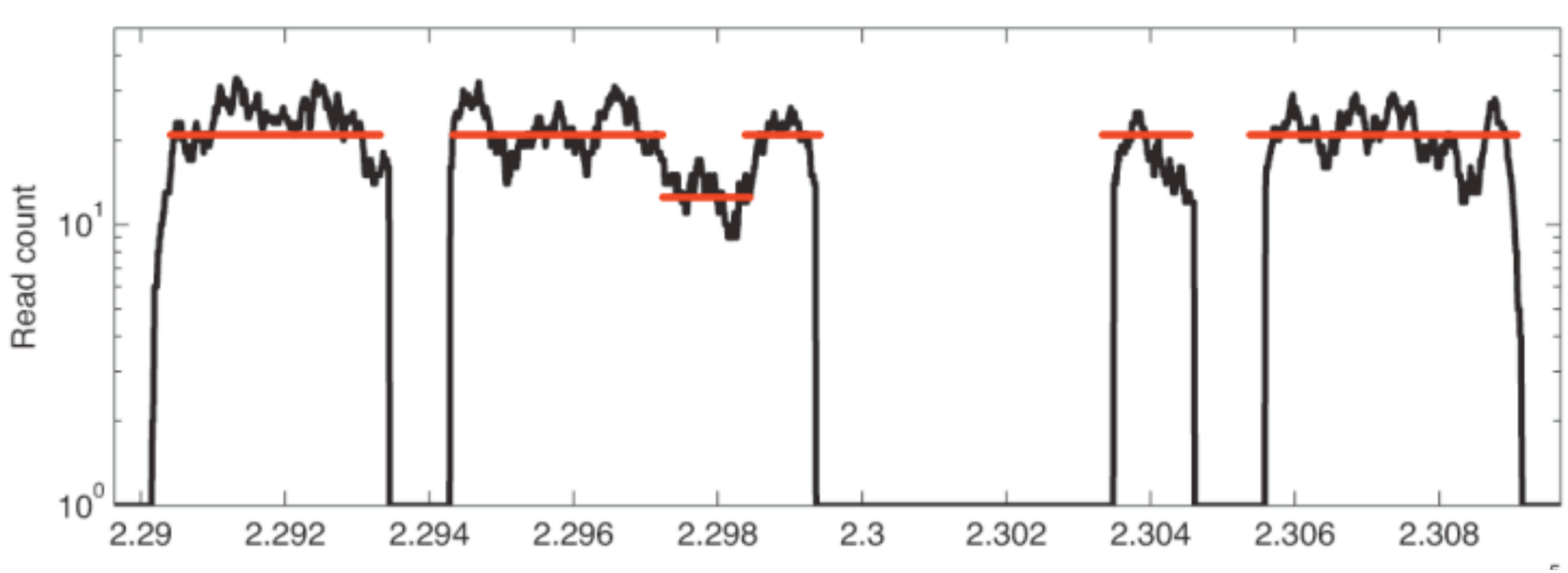


Not-so-random priming

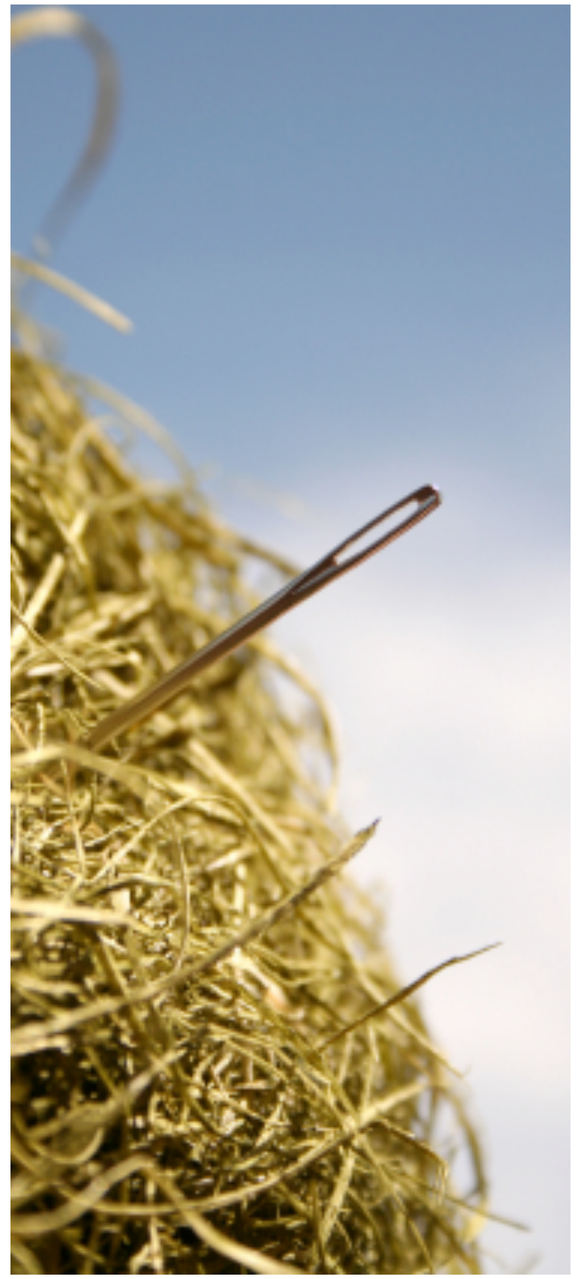


Start/end bias differs by technology

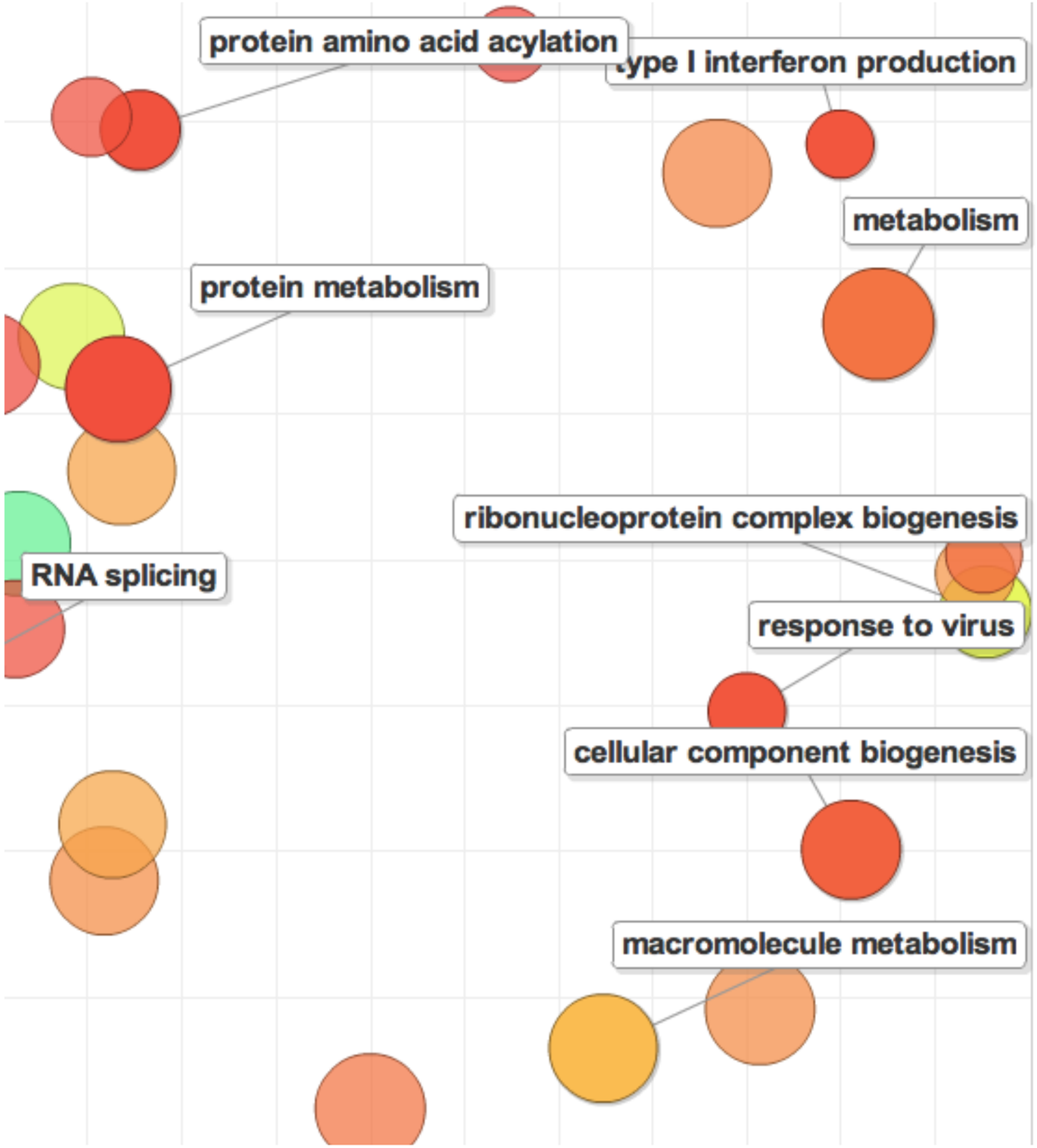

3
1
2
4
5
6
228971
229372
229773
230174
230575

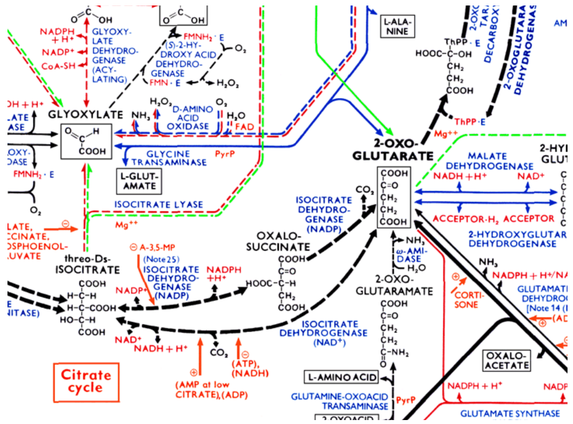

Read count
10^1
10^0
2.29
2.292
2.294
2.296
2.298
2.3
2.302
2.304
2.306
2.308

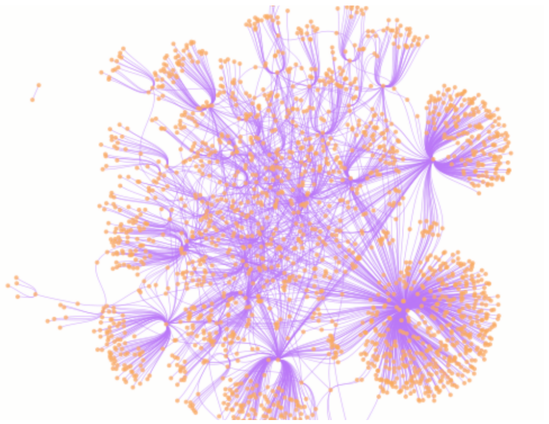
Making sense of your gene list

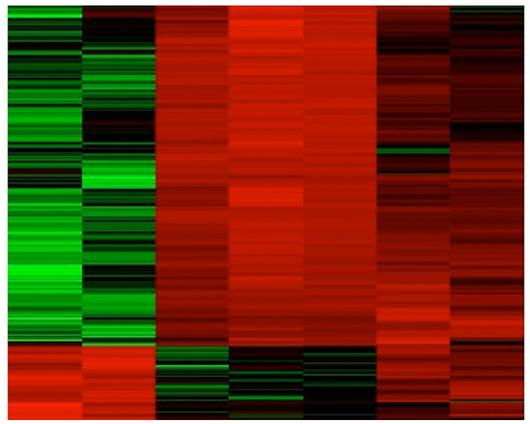


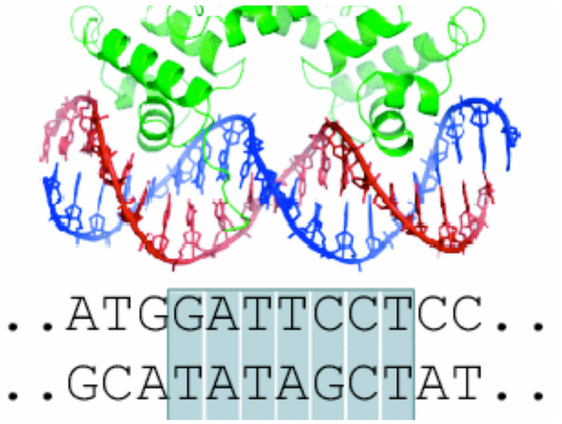
Functional enrichment









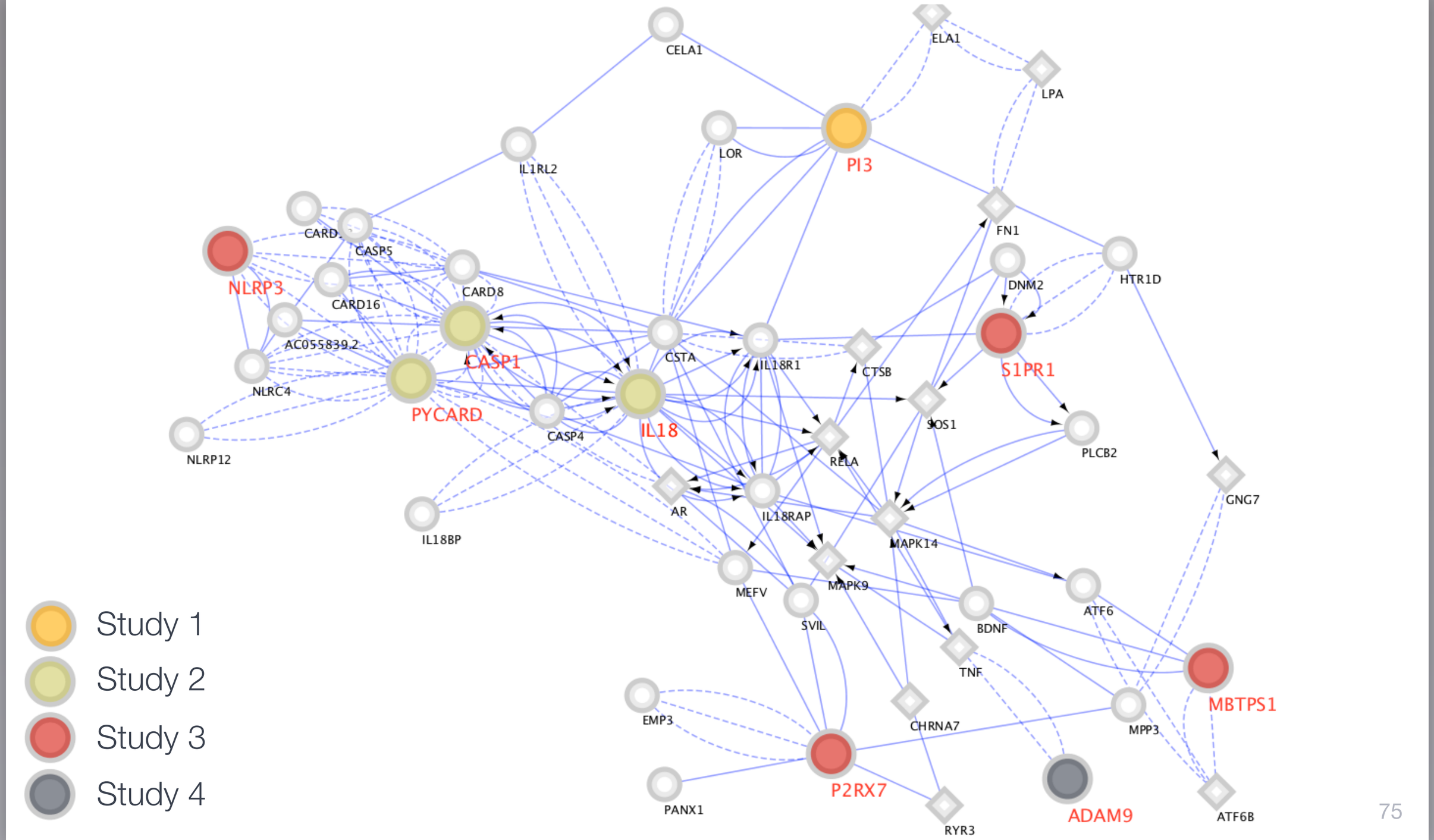


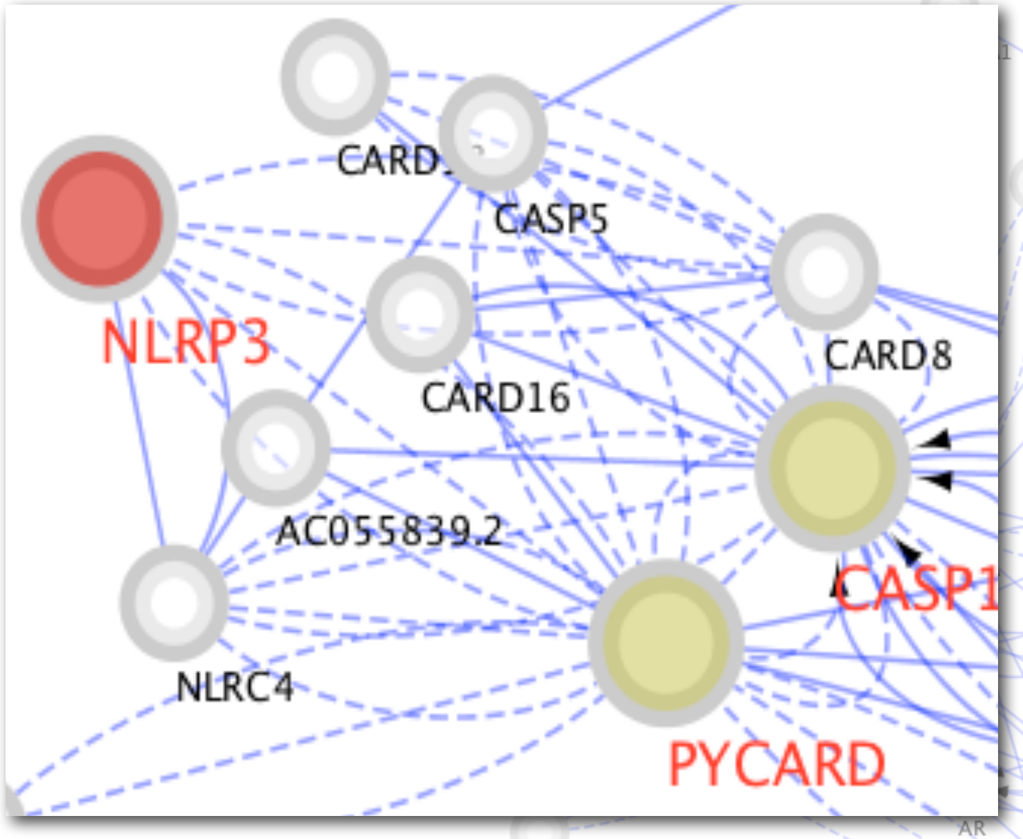
Known interactions

Protein-protein binding

Consistent co-expression

Transcriptional regulation


CELA1
ELA1
LPA
LOR
PI3
IL1RL2
FN1
CARD
CASP5
NLRP3
CARD16
CARD8
DNM2
HTR1D
AC055839.2
CASP1
CSTA
IL18R1
CTSB
S1PR1
NLRC4
PYCARD
SOS1
CASP4
IL18
PLCB2
NLRP12
RELA
GNG7
AR
IL18RAP
IL18BP
MAPK14
MEFV
MAPK9
ATF6
SVIL
BDNF
TNF
MBTPS1
EMP3
CHRNA7
MPP3
P2RX7
PANX1
RYR3
ADAM9
ATF6B
Study 1
Study 2
Study 3
Study 4

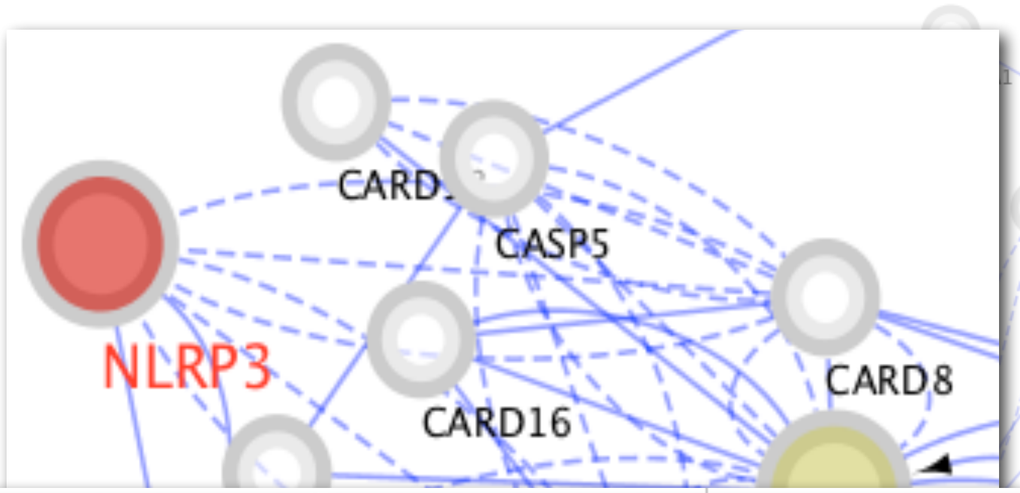


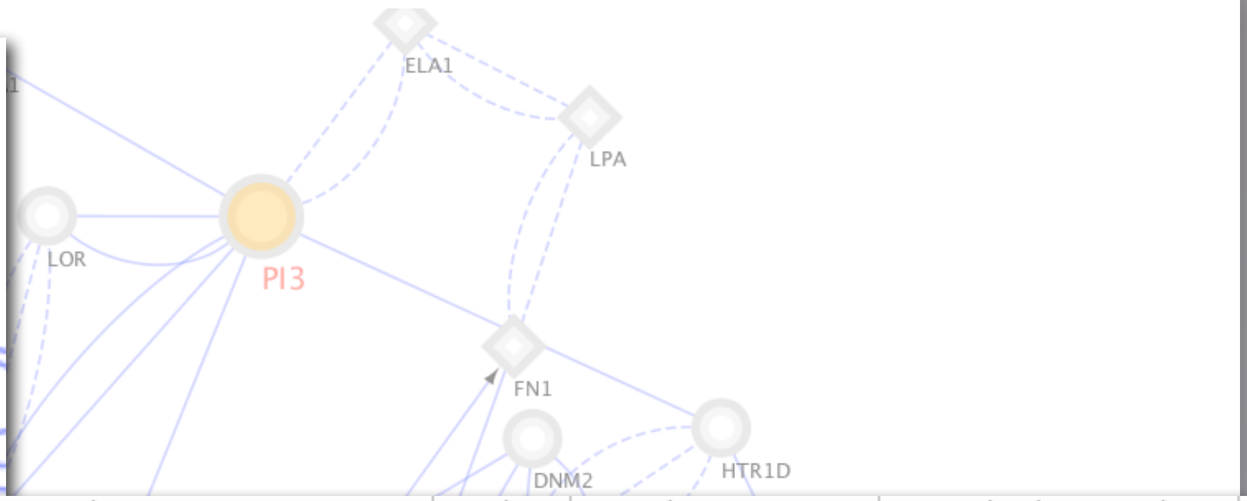
Study 1

Study 2

Study 3

Study 4



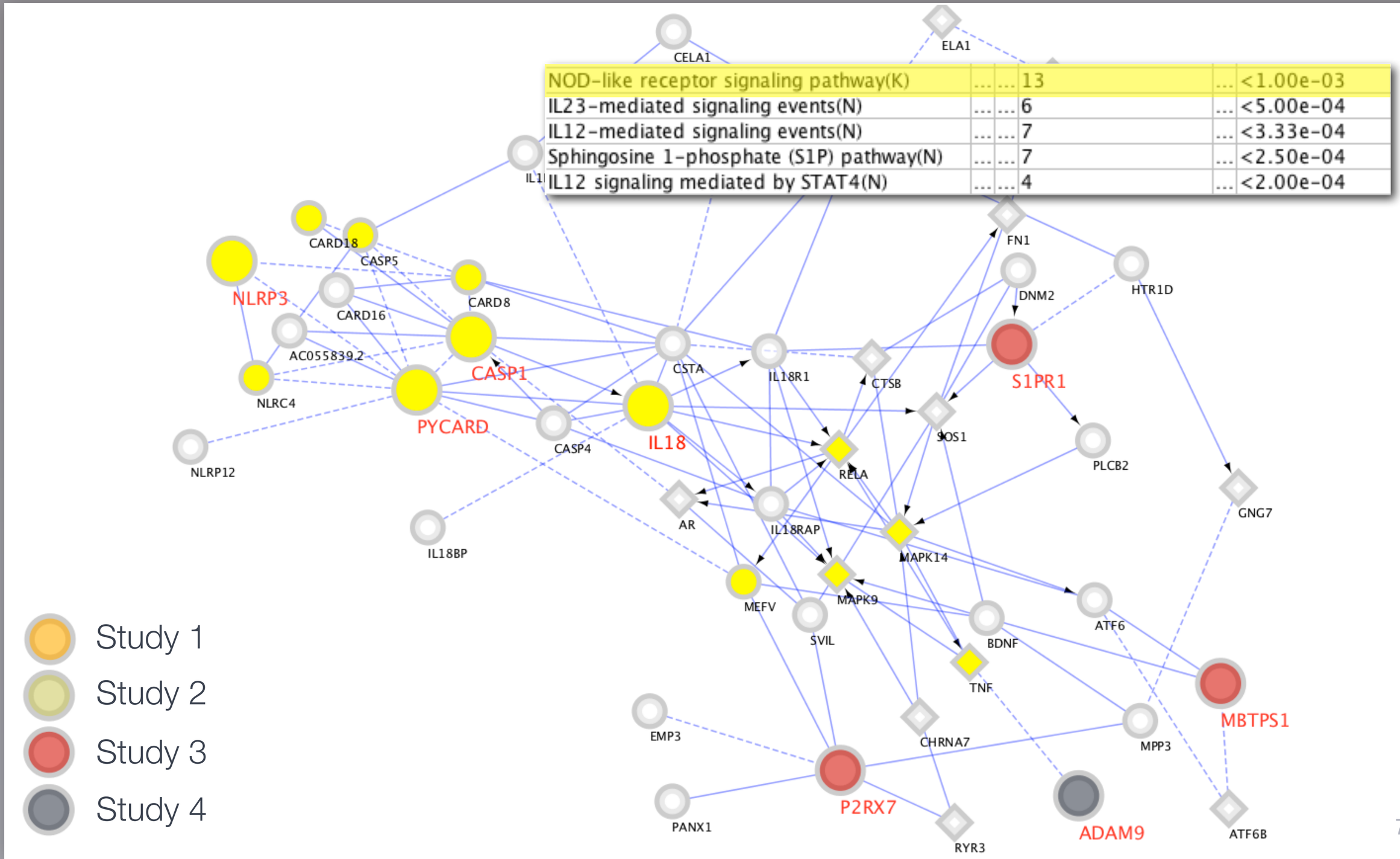


ID	networks	rank	data type	normalized max weight
CASP4 (Co expression H sapiens 2) PYCARD	[Hannenhalli-Cappola-2006]	0	Co-expression	3.365162984771238E...
IL18R1 (Co expression H sapiens 2) CARD8	[Wang-Maris-2006]	0	Co-expression	0.003545847771278...
CASP1 (Pathway H sapiens 2) IL18	[PATHWAYCOMMONS-NCI_NATURE]	2	Pathway	0.03513637205764857
IL18 (Co expression H sapiens 2) PYCARD	[Ross-Perou-2001]	0	Co-expression	0.004547999700193...
CASP5 (Co expression H sapiens 2) IL1RL2	[Wang-Maris-2006]	0	Co-expression	0.001120727970863...
IL18BP (Physical interactions H sapiens 2) IL18	[PATHWAYCOMMONS]	1	Physical interactions	0.05554565914902503
CASP4 (Pathway H sapiens 2) CASP1	[PATHWAYCOMMONS-NCI_NATURE]	2	Pathway	0.17224503150392498
CARD16 (Co expression H sapiens 2) CASP1	[Zangrando-Basso-2009]	0	Co-expression	8.76819811945632E-4
IL1RL2 (Physical interactions H sapiens 2) IL18	[PATHWAYCOMMONS]	1	Physical interactions	0.11139493885906686
CARD16 (Physical interactions H sapiens 2) CARD8	[PATHWAYCOMMONS]	1	Physical interactions	0.06700058824264392
IL18R1 (Pathway H sapiens 2) IL18	[PATHWAYCOMMONS-NCI_NATURE]	2	Pathway	0.020903080601769...
CASP4 (Physical interactions H sapiens 2) IL18	[PATHWAYCOMMONS]	1	Physical interactions	0.014599488029586...
CARD18 (Physical interactions H sapiens 2) CARD8	[PATHWAYCOMMONS]	1	Physical interactions	0.08697183339241087
CASP5 (Physical interactions H sapiens 2) CASP1	[PATHWAYCOMMONS]	1	Physical interactions	0.015267667745317...
CSTA (Co expression H sapiens 2) CASP1	[Ross-Perou-2001, Hannenhalli-Cappola-2006]	0	Co-expression	0.002196498810603...
CARD16 (Co expression H sapiens 2) PYCARD	[Zangrando-Basso-2009]	0	Co-expression	5.060063714509976E...
CASP1 (Co expression H sapiens 2) IL18	[Wang-Maris-2006]	0	Co-expression	0.001987161993368...
CSTA (Co expression H sapiens 2) IL18	[Wang-Maris-2006, Zangrando-Basso-2009, Burcz...	0	Co-expression	0.003217535435700...

Study 3

Study 4

NOD-like receptor signaling pathway(K)	...	...	13	...	<1.00e-03
IL23-mediated signaling events(N)	...	...	6	...	<5.00e-04
IL12-mediated signaling events(N)	...	...	7	...	<3.33e-04
Sphingosine 1-phosphate (S1P) pathway(N)	...	...	7	...	<2.50e-04
IL12 signaling mediated by STAT4(N)	...	...	4	...	<2.00e-04



Next steps

Afgan et al. BMC Bioinformatics 2010, 11(Suppl 12):S4
http://www.biomedcentral.com/1471-2105/11/S12/S4

BMC Bioinformatics

PROCEEDINGS Open Access

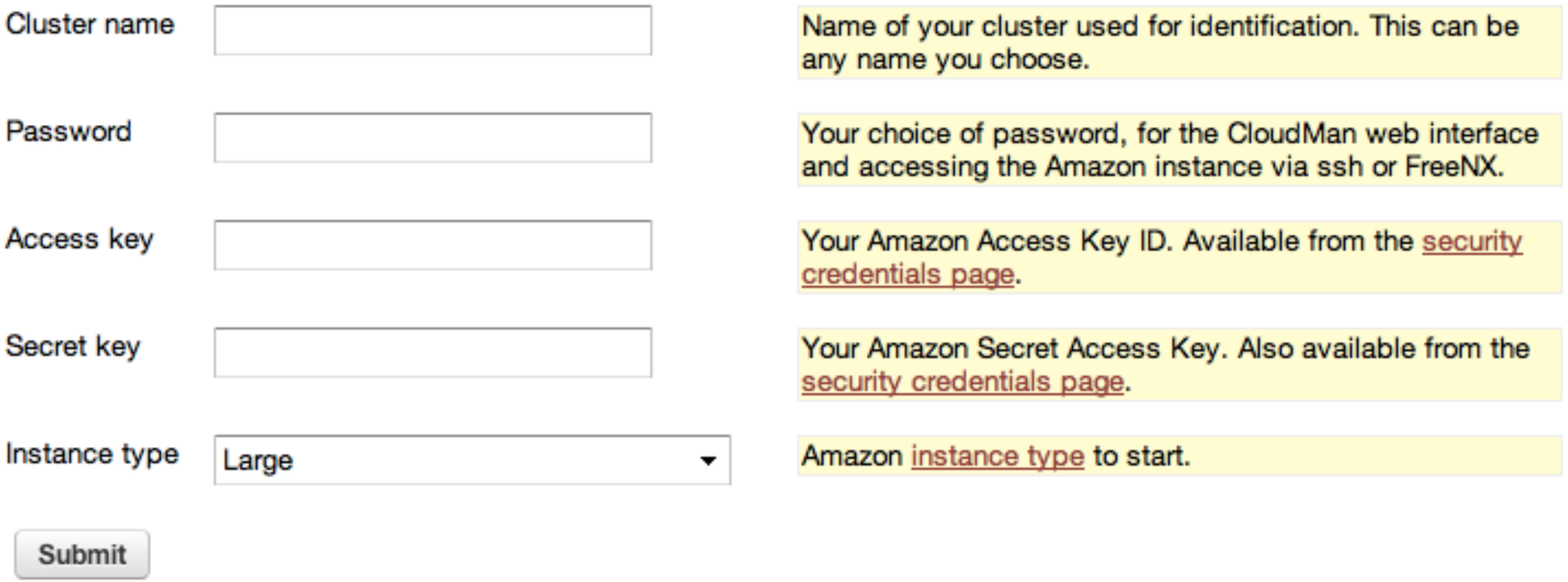
Galaxy CloudMan: delivering cloud compute clusters

Enis Afgan[1], Dannon Baker[1], Nate Coraor[2], Brad Chapman[3], Anton Nekrutenko[2], James Taylor[1*]

From The 11th Annual Bioinformatics Open Source Conference (BOSC) 2010
Boston, MA, USA. 9-10 July 2010

Run your own Galaxy instance

Instant Galaxy cluster



BioCloudCentral

https://biocloudcentral.herokuapp.com/launch

Need for Unix

http://ritg.med.harvard.edu/classes.html

```
2479749 -rw-r--r--   1 root  wheel       53B 23 Jun  2009 networks
2476106 -r--r--r--   1 root  wheel      1.6K 18 May  2009 newsyslog.conf
2377284 drwxr-xr-x   4 root  wheel      136B  3 Sep  2009 newsyslog.d
2387394 -rw-r--r--   1 root  wheel      132B 10 Jul  2009 notify.conf
2483297 -rw-r--r--   1 root  wheel      366B 18 May  2009 ntp-restrict.conf
2733572 -rw-r--r--@  1 root  wheel       27B 20 Oct  2009 ntp.conf
2471151 drwxr-xr-x   7 root  wheel      238B 17 Jul  2009 openldap
4929956 drwxr-xr-x   3 root  wheel      102B 30 Nov  2007 opt
2403948 drwxr-xr-x  17 root  wheel      578B 23 Dec  2010 pam.d
2386884 -rw-r--r--   1 root  wheel      3.6K 23 Jun  2009 passwd
2479751 -rw-r--r--   1 root  wheel       45B 23 Jun  2009 paths
2475307 drwxr-xr-x   3 root  wheel      102B 11 Jul  2009 paths.d
2475966 -rw-r--r--   1 root  wheel      1.2K 19 Jul  2009 pear.conf
2466447 drwxr-xr-x   5 root  wheel      170B 18 May  2009 periodic
13701882 -r--r--r--   1 root  wheel       67K  9 Mar 00:49 php.ini.default
2733641 -r--r--r--   1 root  wheel       44K  6 Feb  2009 php.ini.default-5
2485974 drwxr-xr-x  23 root  wheel      782B 27 Jun  2010 postfix
2486265 drwxr-xr-x   2 root  wheel       68B  1 Aug  2009 ppp
2476001 -r--r--r--   1 root  wheel      189B  4 May  2009 profile
2386885 -rw-r--r--   1 root  wheel      5.6K 23 Jun  2009 protocols
2444032 drwxr-xr-x   4 root  wheel      136B  3 Sep  2009 racoon
2387062 -rw-r--r--   1 root  wheel      1.6K 25 Jul  2009 rc.common
2444075 -rw-r--r--   1 root  wheel      5.0K 25 Jul  2009 rc.netboot
2442640 lrwxr-xr-x   1 root  wheel       20B  3 Sep  2009 resolv.conf -> /\
lv.conf
2479752 -rw-r--r--   1 root  wheel        0B 23 Jun  2009 rmtab
2386886 -rw-r--r--   1 root  wheel      971B 23 Jun  2009 rpc
2483260 -rw-r--r--   1 root  wheel      983B 15 Jul  2009 rtadvd.conf
2403837 drwxr-xr-x   7 root  wheel      238B 16 Jun  2009 security
2386887 -rw-r--r--   1 root  wheel      662K 23 Jun  2009 services
2386888 -rw-r--r--   1 root  wheel      179B 23 Jun  2009 shells
4613036 -rw-r--r--   1 root  wheel      2.9K 27 Jun  2010 smb.conf
2497382 -rw-r--r--   1 root  wheel      2.9K 22 May  2009 smb.conf.old
4599486 -rw-r--r--   1 root  wheel      2.9K  6 May  2010 smb.conf.template
2482910 drwxr-xr-x   4 root  wheel      136B 28 Jul  2009 snmp
2403959 -rw-r--r--   1 root  wheel      1.5K 11 Jul  2009 ssh_config
2403960 -rw-r--r--   1 root  wheel      3.6K 11 Jul  2009 sshd_config
2497815 -r--r-----   1 root  wheel      1.2K 23 Jun  2009 sudoers
2386889 -rw-r--r--   1 root  wheel      772B 23 Jun  2009 syslog.conf
2386890 -rw-r--r--   1 root  wheel      1.4K 23 Jun  2009 ttys
2475311 drwxr-xr-x   4 root  wheel      136B 28 Jul  2009 xgrid
2479753 -rw-r--r--   1 root  wheel        0B 23 Jun  2009 xtab
2503639 -r--r--r--   1 root  wheel      126B 11 May  2009 zshenv
Ohos-MacBook-Air:etc oho$
```

software carpentry

Bootcamps Lessons Blog FAQ About Search

TEACHING LAB SKILLS FOR

SCIENTIFIC COMPUTING

Who We Are

Our volunteers teach basic software skills to researchers in science, engineering, and medicine. Founded in 1998, we are now part of the Mozilla Science Lab.

What We Do

We run bootcamps all over the world, and provide open access material for self-paced instruction. We also run a training program for people who'd like to help us teach.

How To Help

Like all volunteer organizations, we depend on you to help us help others. You can host a bootcamp, help create new teaching materials, or improve the tools we use.

Software Carpentry

bcbio-nextgen.readthedocs.org

Not everything as complicated...

- ... but sequencing is becoming a commodity

Talk to us early

Involvement in study design to optimize experiments

Contact

ohofmann@hsph.harvard.edu

@fiamh

http://compbio.sph.harvard.edu/chb/