Extracting relevant information from UHTS data: analysis pipelines (smallRNA)

Patricia Otten

3th July 2012

JOBIM

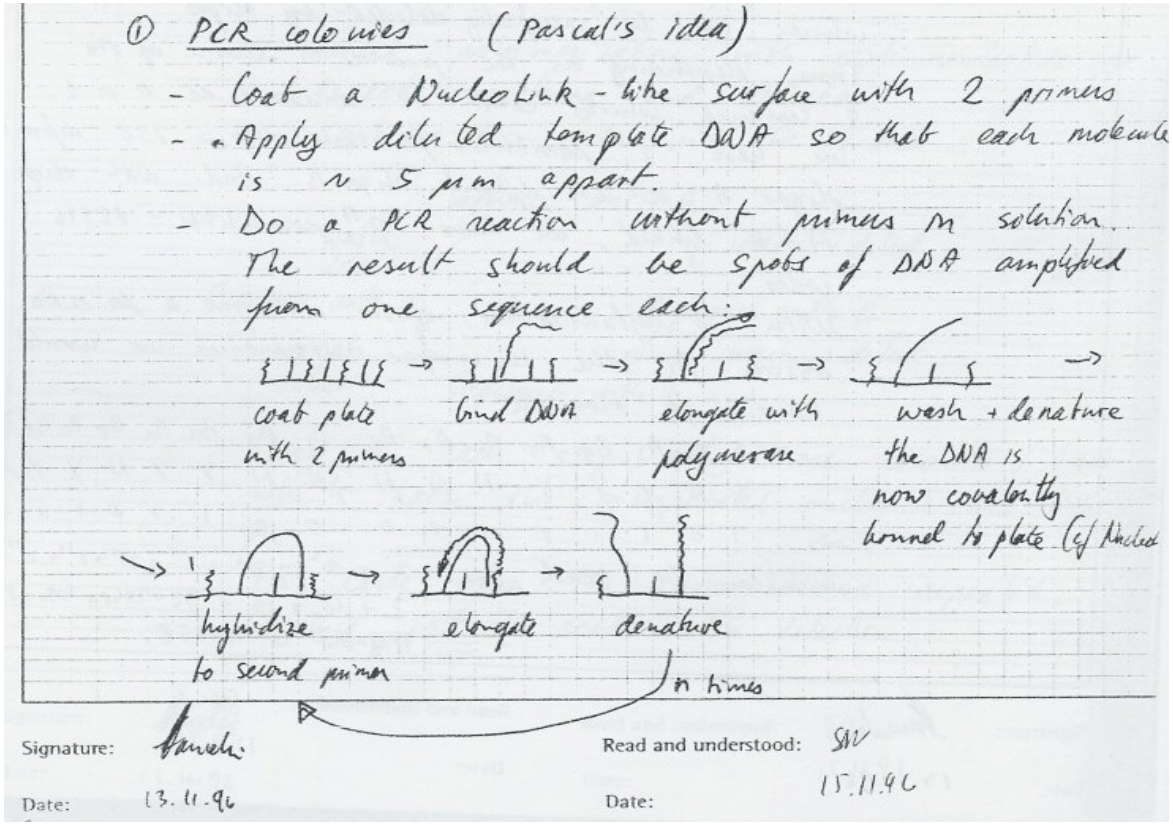
Rennes - France

Fasteris SA: Illumina sequencing

- founded in 2003 by L. FARINELLI and M. OSTERAS
- 2012: about 20 collaborators
- capillary and UHTS sequencing + bioinformatics
- private and academic labs
- no business plan, no external investors, no sales forces

Illumina sequencing

Key technology based on the concept of DNA colonies, invented in 1996 at the GlaxoWellcome's Geneva Biomedical Research Institute



Mayer P., Farinelli L. and Kawashima, E.,
1997, Patent application WO 98/44151

Illumina sequencing: step1

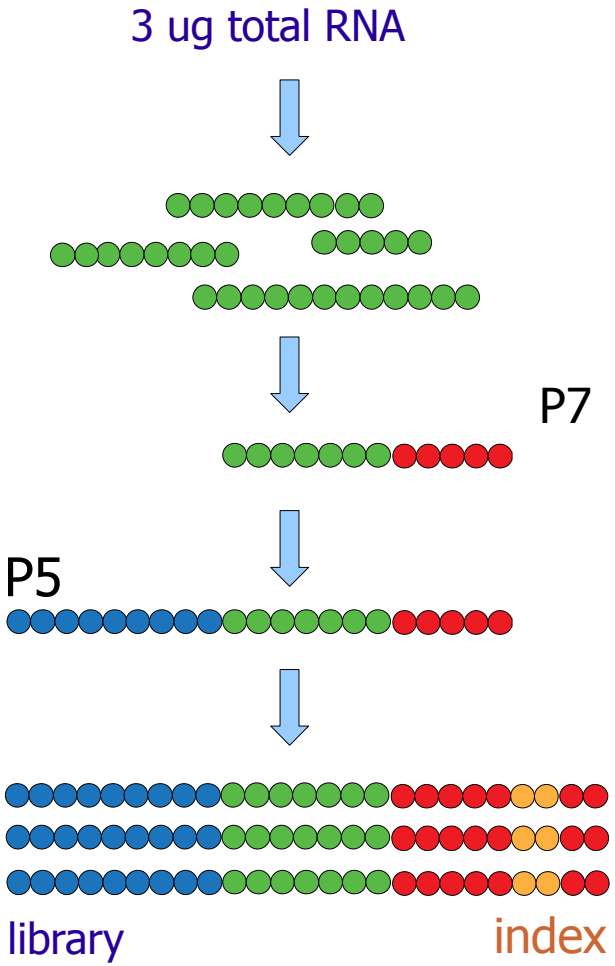
Library preparation (smallRNA protocol)

selection of small RNAs (20-30 nt) acrylamide gel purification

single-stranded ligation of the 3' adapter

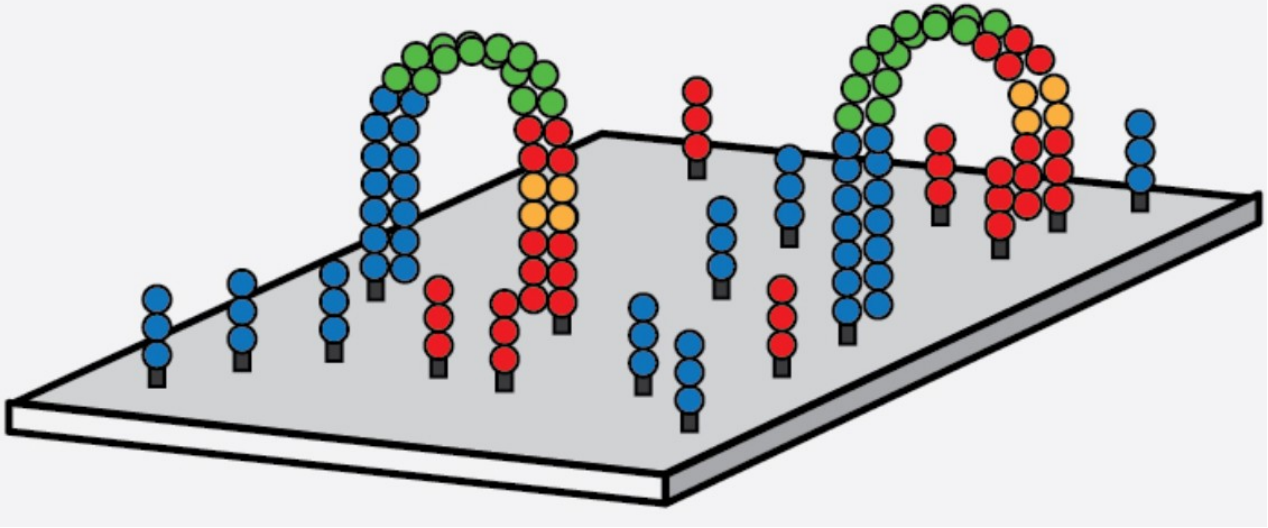
single-stranded ligation of the 5' adapter

reverse transcription, PCR, index addition, gel purification



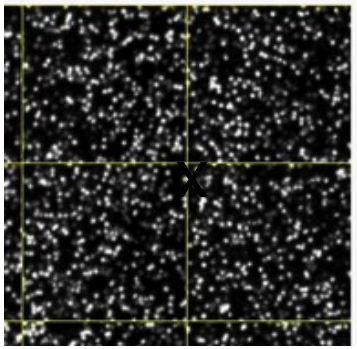
Illumina sequencing: step2

Flowcell preparation

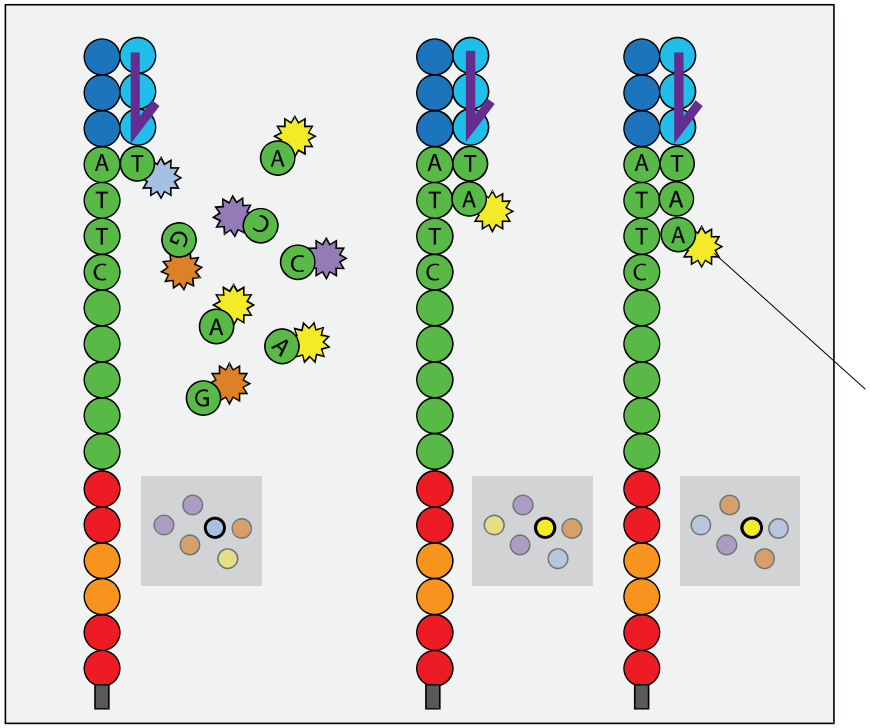


Templates are hybridized to a surface (flowcell) and in situ amplified (bridge amplification) to form DNA colonies.

- each colony produces one read
- all colonies are sequenced in parallel
- ~150 mio passed filter reads per lane



Illumina sequencing: step 3



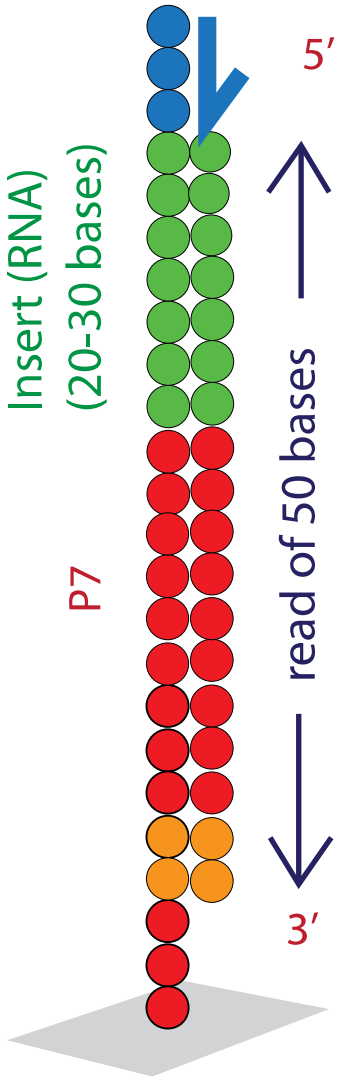
Sequencing

Incorporation of reversible-terminator nucleotides labeled with fluorescent dyes

- base per base sequencing (50, 100 cycles, SR or PE)
- laser excitation and image capture; release of dye;
- intensities extraction and base calling by RTA software

1x100 run: 1 week; 1.5 TB intensities; 200 GB sequences;

Trimming (smallRNAs)



Adapter trimming

5′ ← read of 50 bases → 3′

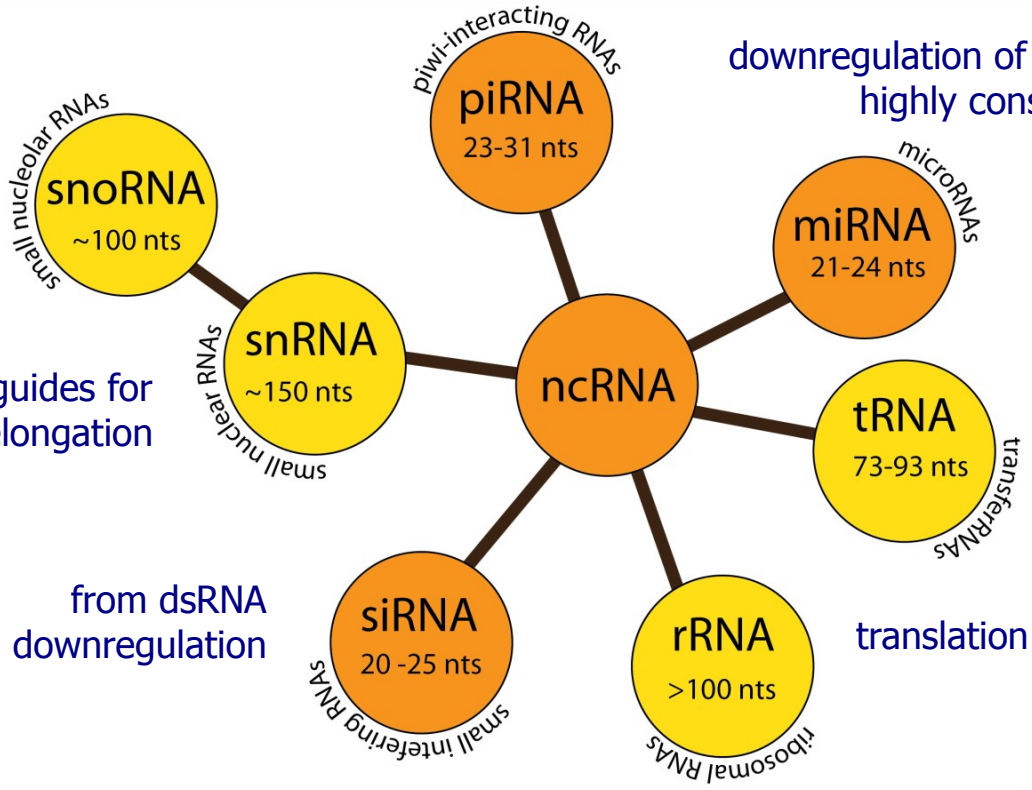
AAGGTGATTGTGGCTTGGAATTCTCGGG
AGAAGGTGATTGTGGCTTGGAATTCTCG
GTGTGTGTGTGAGTGTGTTGGAATTCTC
AGAAGGTGATTGTGGCTTGGAATTCTCG
CTAGGTGATGAGTCATGGAATTCTCGGG
GAATGGTAGAACTCACACTTGGAATTCT
TTCTGTGATAACTGAATGGAATTCTCGG
GCATGGTAGAACTCACACTTGGAATTCT
CAGAGGTGAGTGTGGCTTGGAATTCTCG

← inserts →

Introduction to smallRNAs

transposons silencing
poorly conserved

chemical modifications of other RNAs, mainly rRNAs, tRNAs and snRNAs



downregulation of genes
highly conserved

RNA splicing, guides for telomere elongation

from dsRNA downregulation

translation

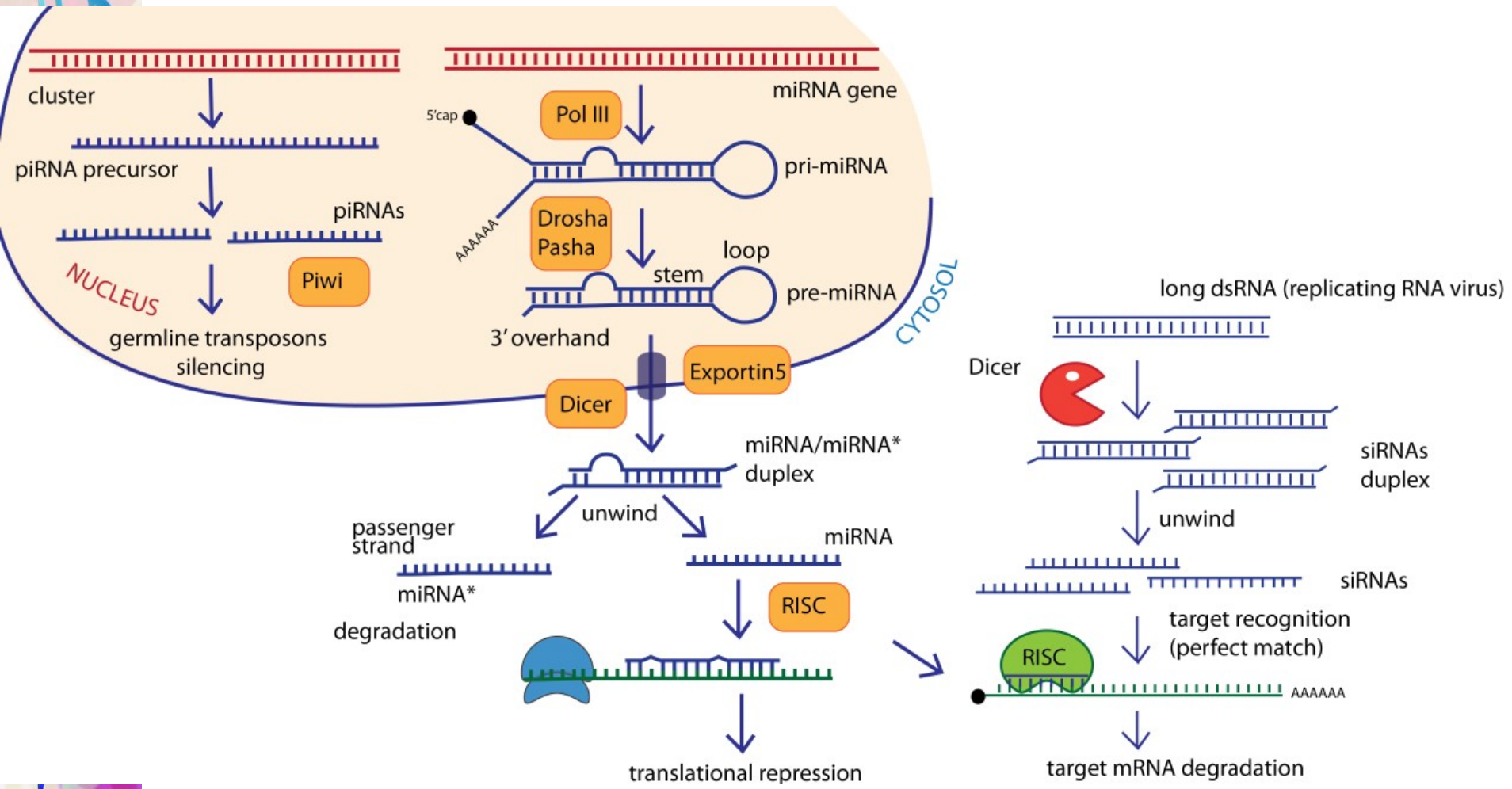
Expression analysis

Virus assembly

Sequencing by siRNA: a novel generic tool for virus discovery

Kreuze et al. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. ***Virology*** **388: 1-7**

Introduction to smallRNAs



Pipelines and automation

produce meaningfull data

time resources

www.photo-dictionary.com

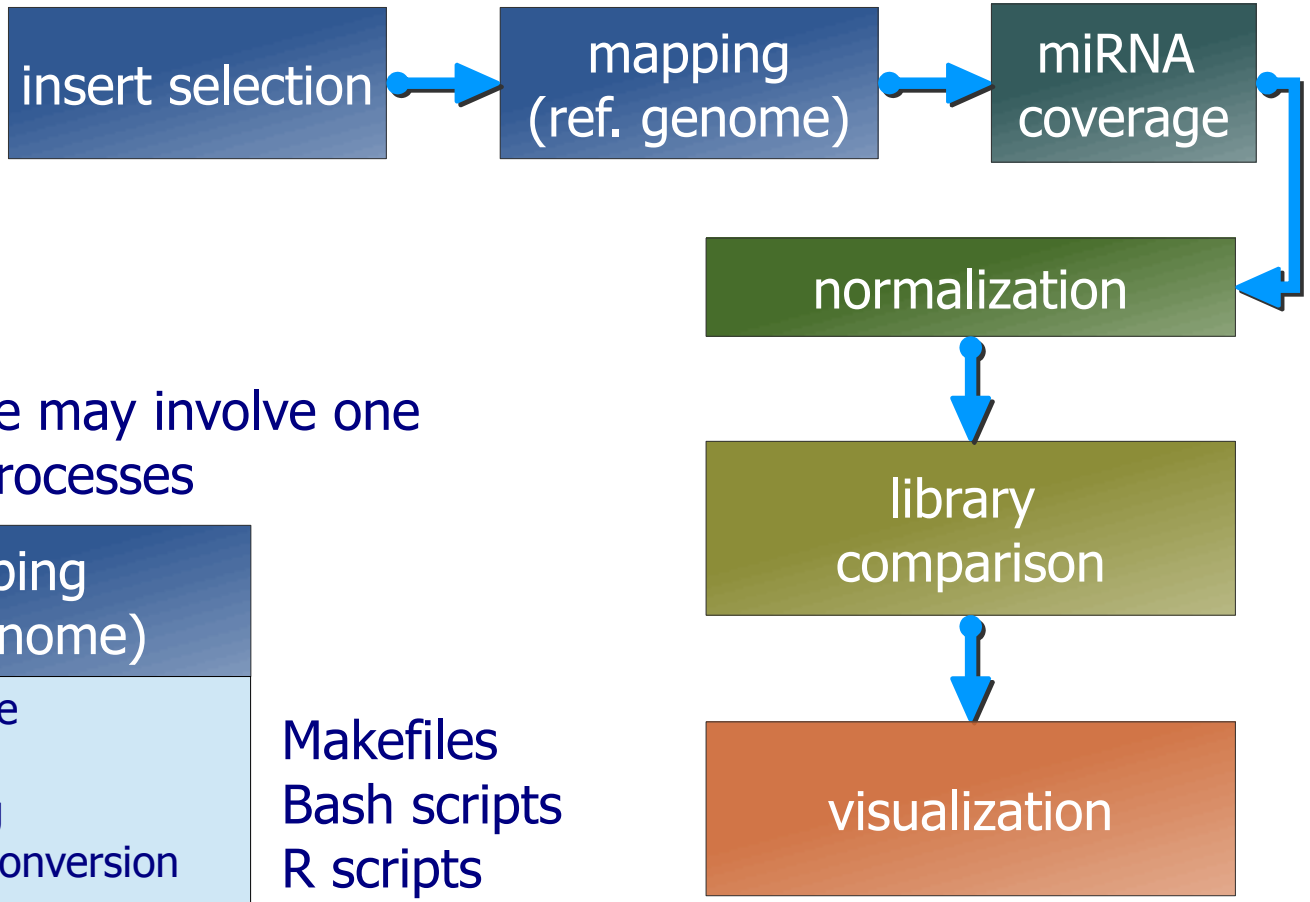
→ automation + checks

→ handle unexpected issues, keep time for the client

→ a pipeline is a set of predetermined tasks that have to be executed to complete a specific analysis

Pipelines and automation

Eg: comparison of libraries in terms of miRNA coverage



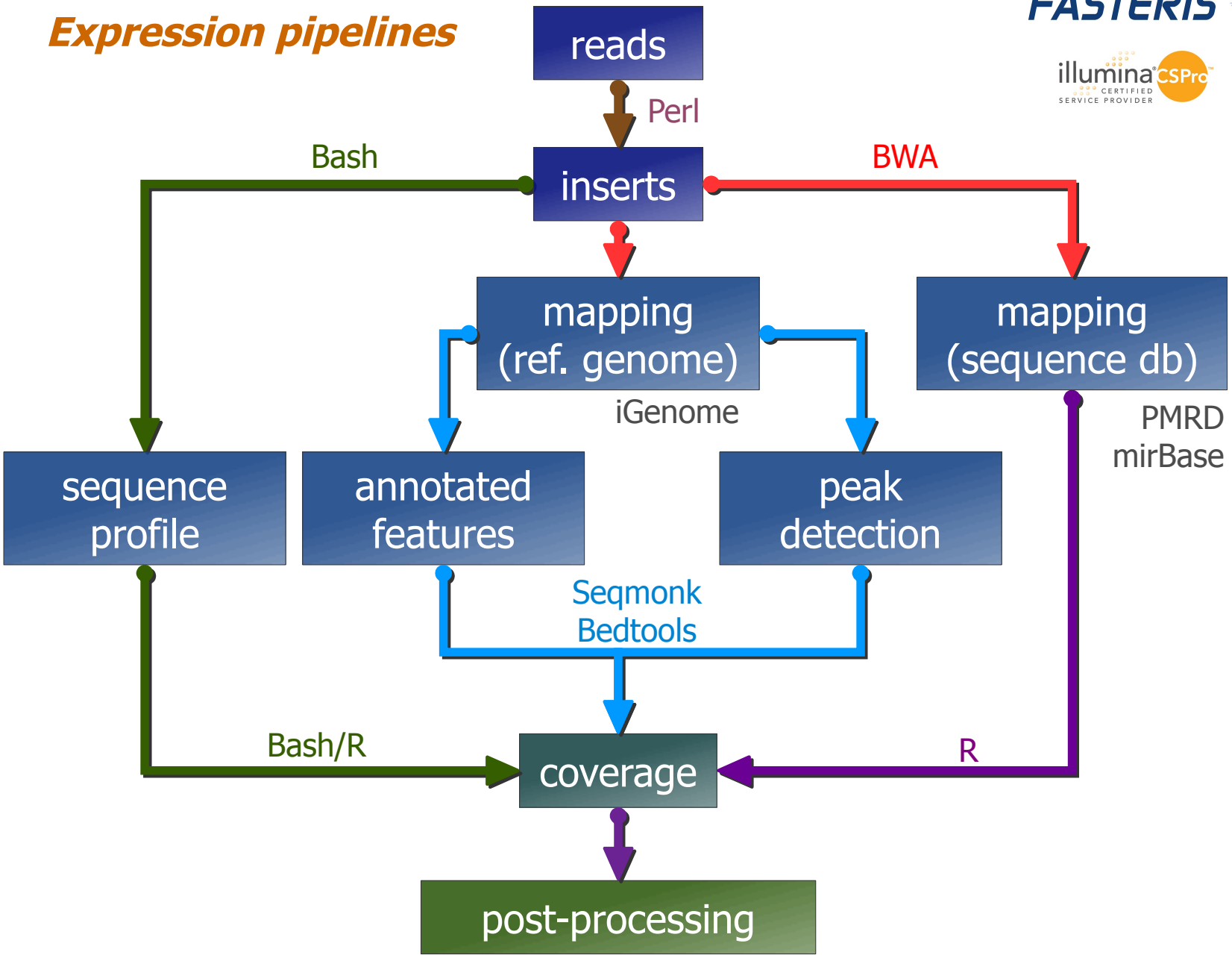
Each module may involve one or several processes

mapping (ref. genome)

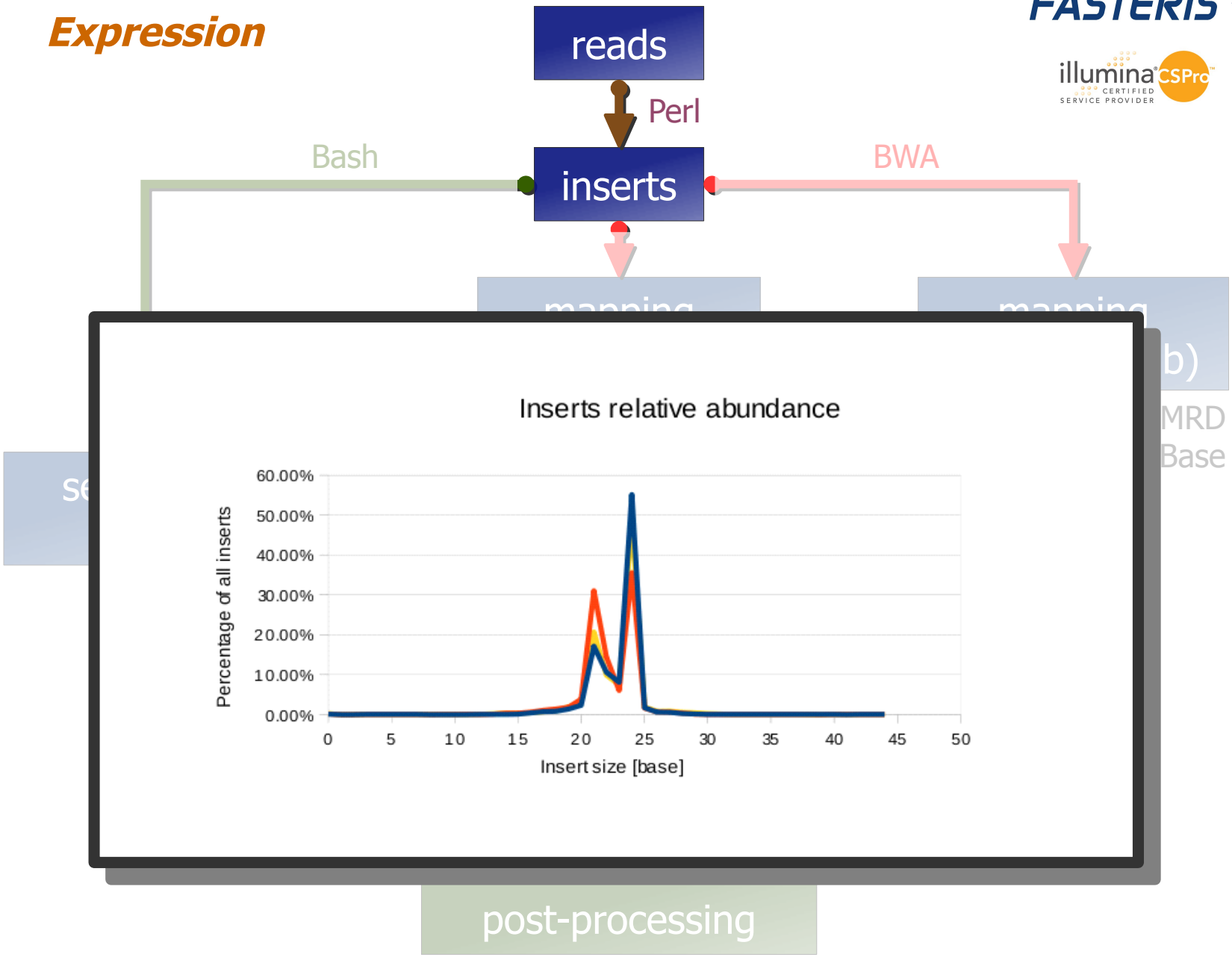
1. reference
2. indexing
3. mapping
4. format conversion
5. reporting

Makefiles
Bash scripts
R scripts

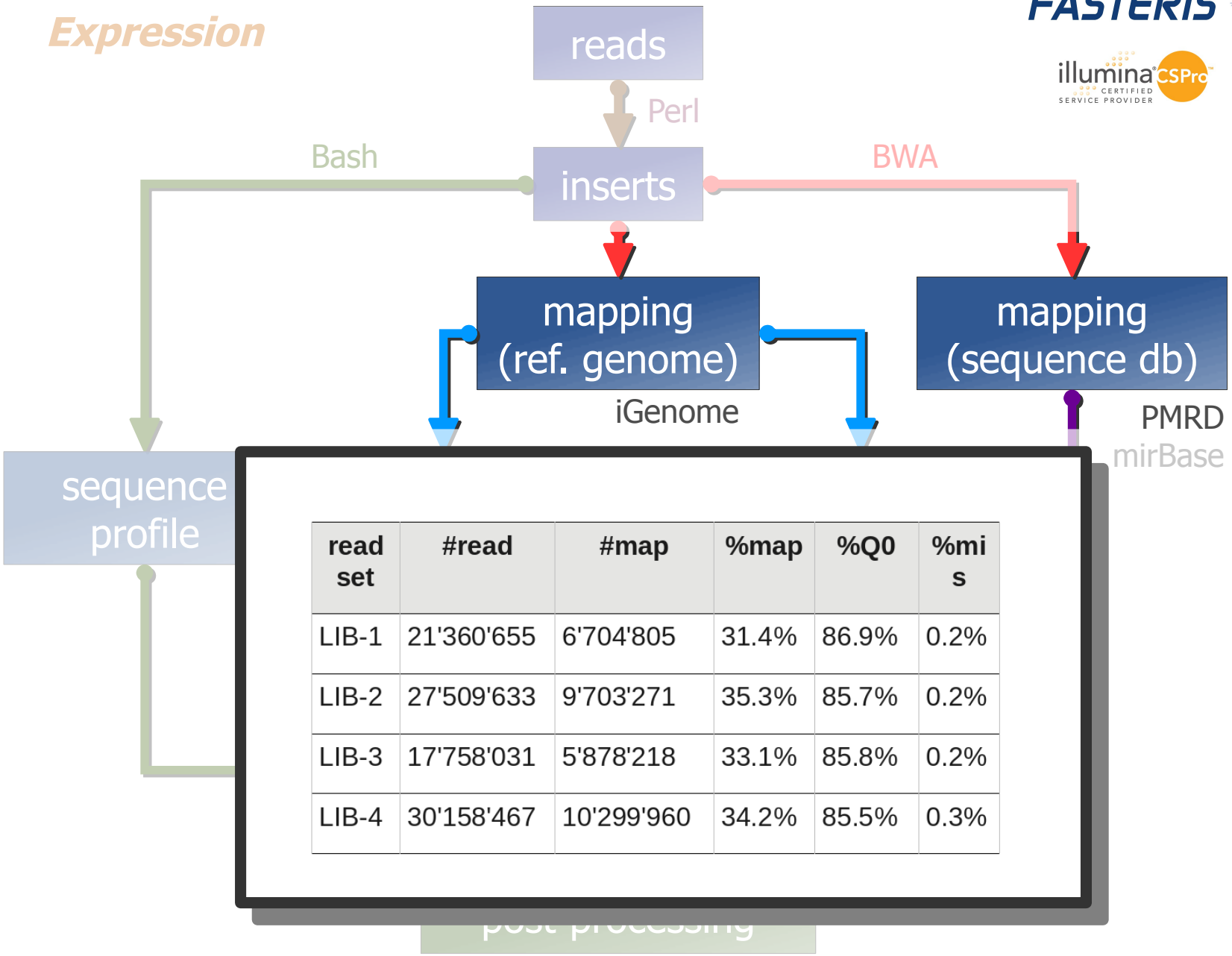
Expression pipelines



Expression

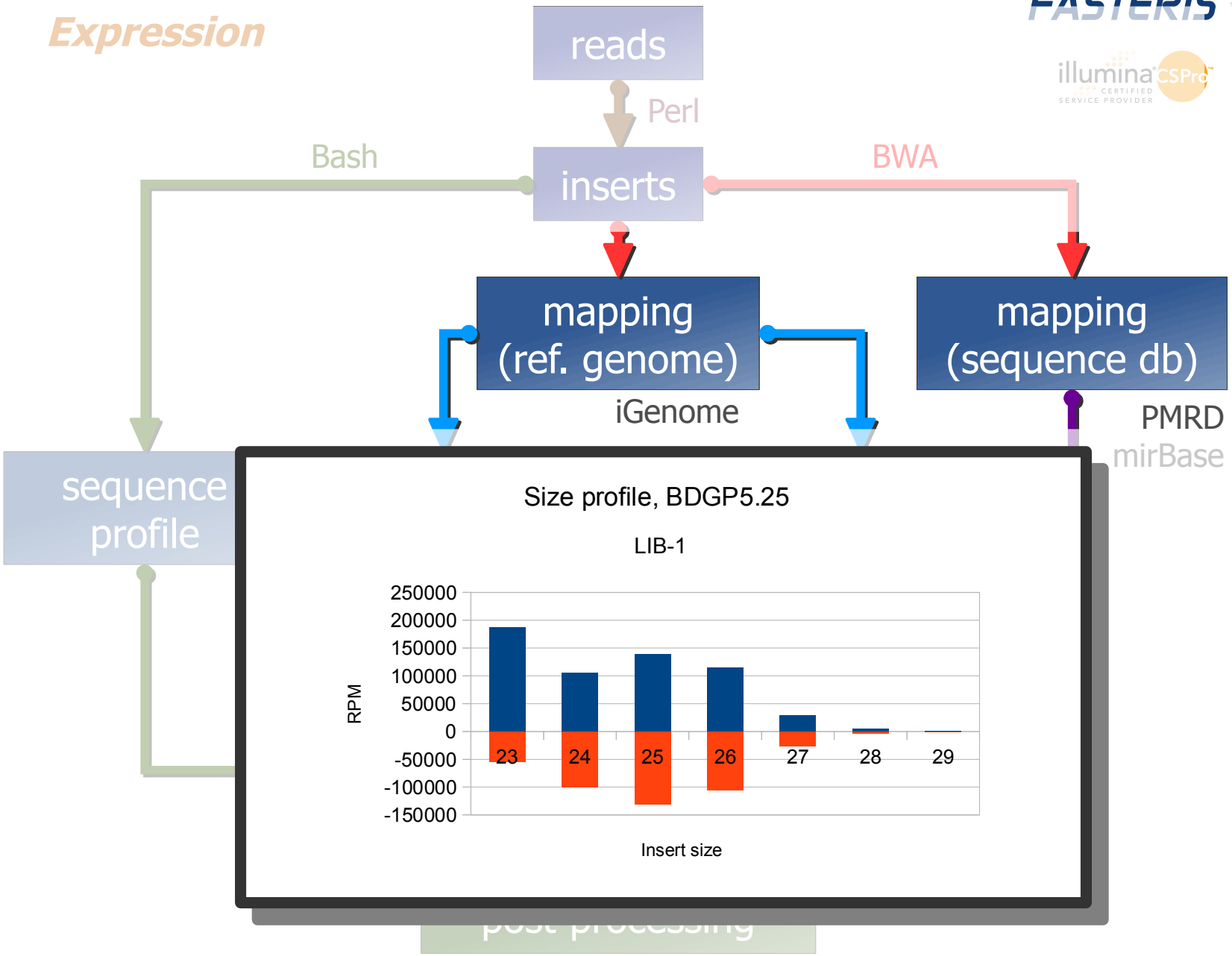

reads
Perl
Bash
BWA
inserts
mapping
mapping
b)
MRD
Base
Inserts relative abundance
Percentage of all inserts
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
5
10
15
20
25
30
35
40
45
50
Insert size [base]
post-processing

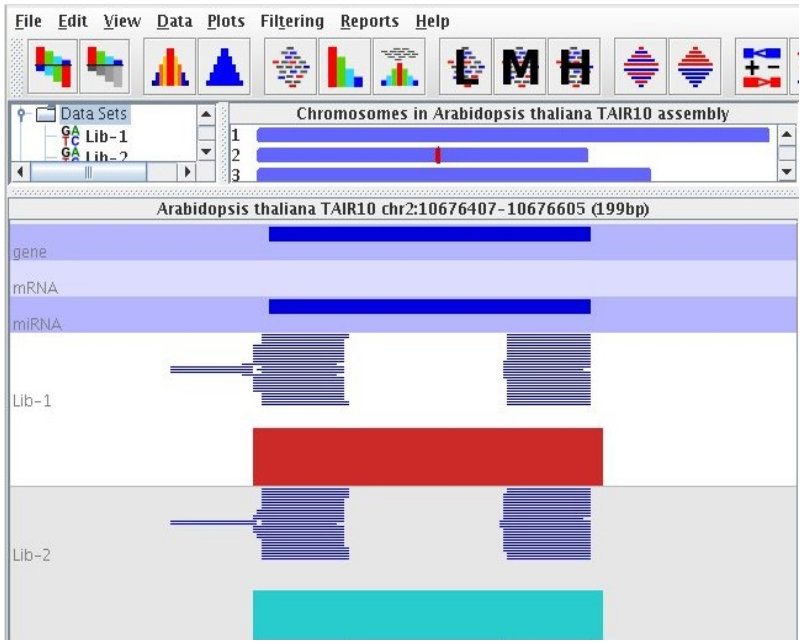
Expression

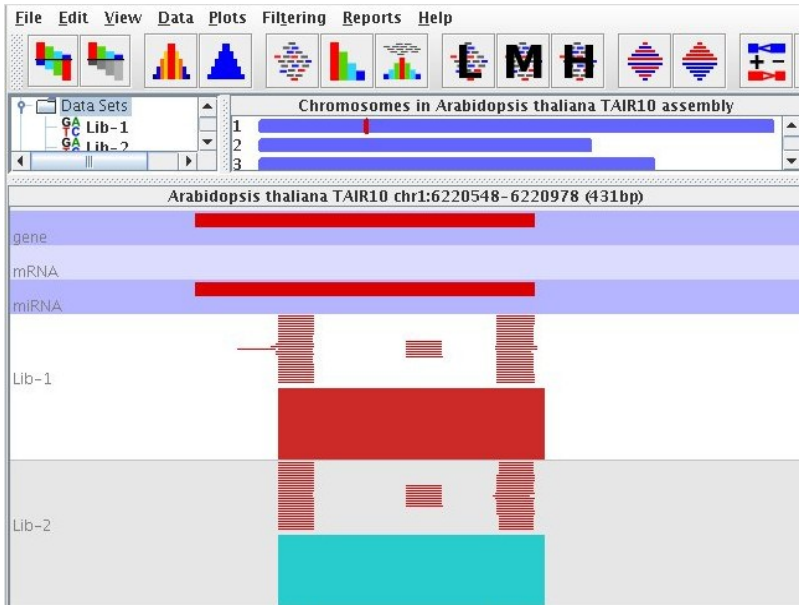


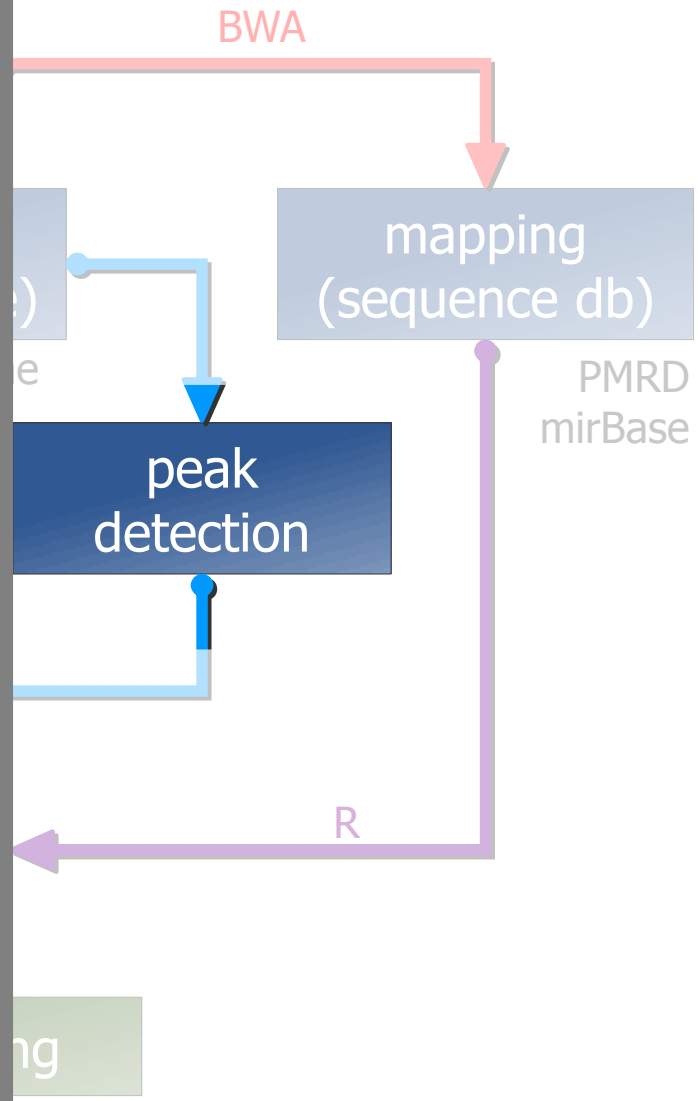
read set	#read	#map	%map	%Q0	%mis
LIB-1	21'360'655	6'704'805	31.4%	86.9%	0.2%
LIB-2	27'509'633	9'703'271	35.3%	85.7%	0.2%
LIB-3	17'758'031	5'878'218	33.1%	85.8%	0.2%
LIB-4	30'158'467	10'299'960	34.2%	85.5%	0.3%

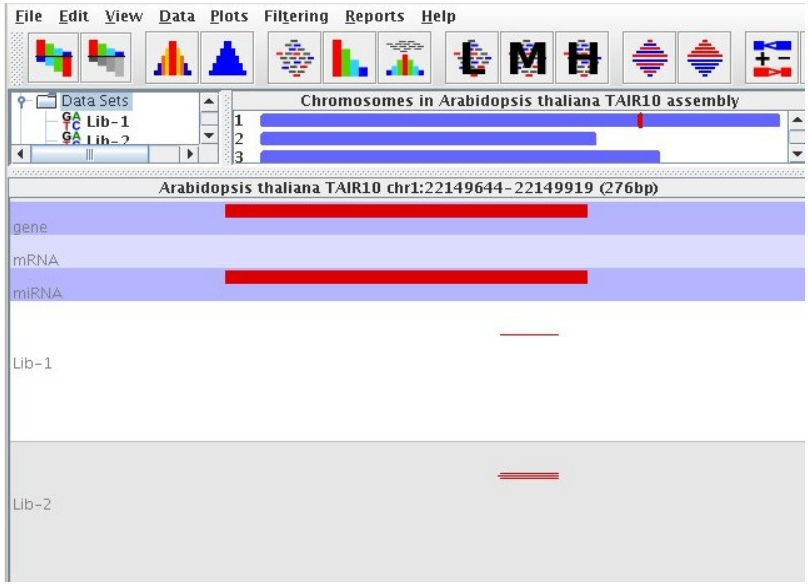
Expression

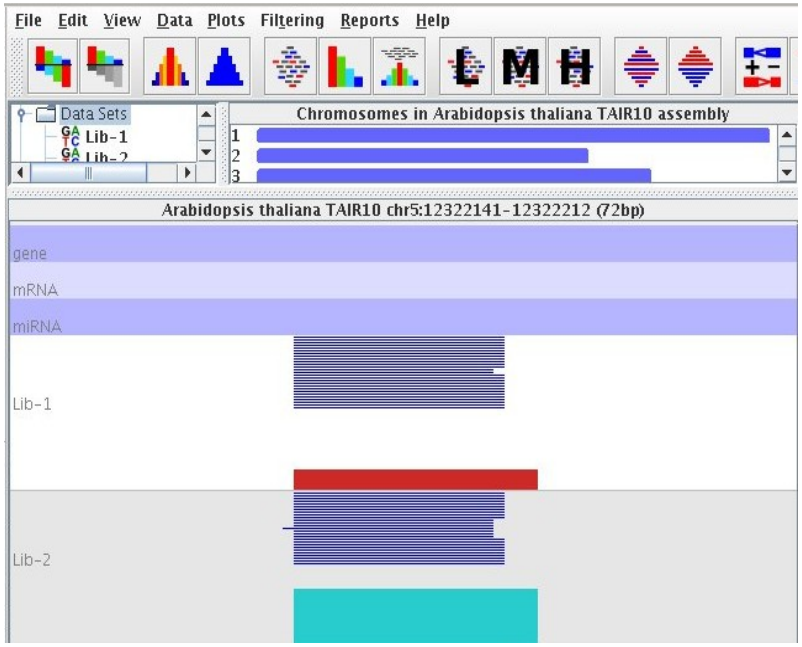

reads
Perl
Bash
inserts
BWA
mapping
(ref. genome)
iGenome
mapping
(sequence db)
PMRD
mirBase
sequence
profile
Size profile, BDGP5.25
LIB-1
250000
200000
150000
100000
50000
0
-50000
-100000
-150000
RPM
23
24
25
26
27
28
29
Insert size

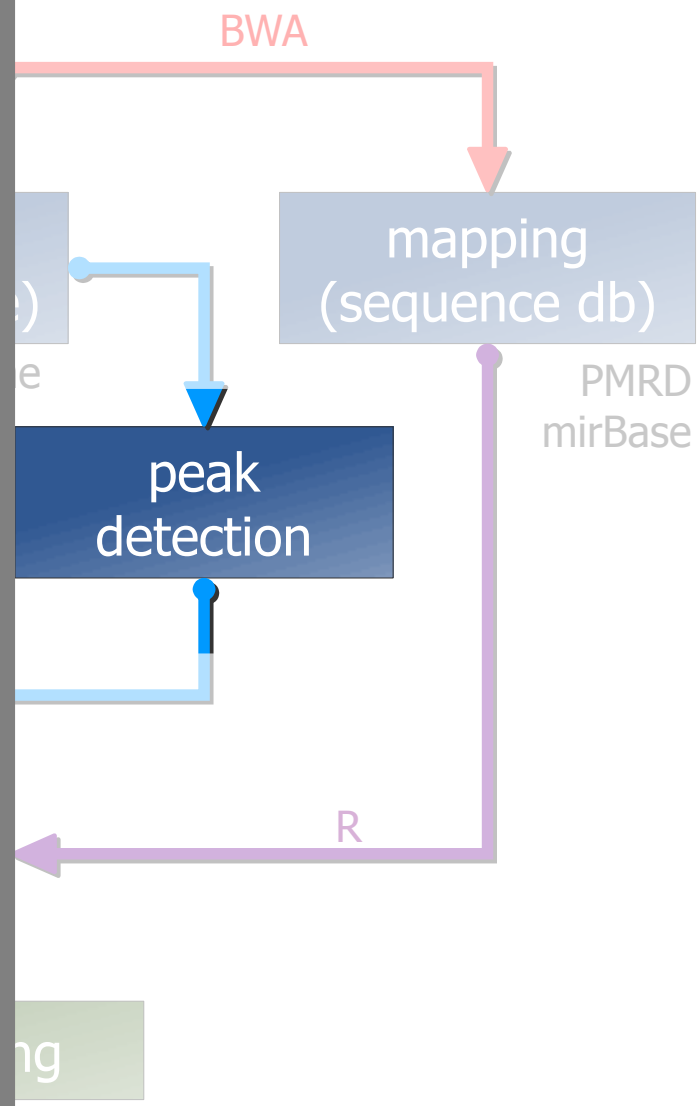

File Edit View Data Plots Filtering Reports Help
Data Sets
Lib-1
Lib-2
Chromosomes in Arabidopsis thaliana TAIR10 assembly
1
2
3
Arabidopsis thaliana TAIR10 chr2:10676407-10676605 (199bp)
gene
mRNA
miRNA
Lib-1
Lib-2


File Edit View Data Plots Filtering Reports Help
Data Sets
Lib-1
Lib-2
Chromosomes in Arabidopsis thaliana TAIR10 assembly
1
2
3
Arabidopsis thaliana TAIR10 chr1:6220548-6220978 (431bp)
gene
mRNA
miRNA
Lib-1
Lib-2


BWA
mapping
(sequence db)
PMRD
mirBase
peak
detection
R


File Edit View Data Plots Filtering Reports Help
Data Sets
Lib-1
Lib-2
Chromosomes in Arabidopsis thaliana TAIR10 assembly
1
2
3
Arabidopsis thaliana TAIR10 chr1:22149644-22149919 (276bp)
gene
mRNA
miRNA
Lib-1
Lib-2


File Edit View Data Plots Filtering Reports Help
Data Sets
Lib-1
Lib-2
Chromosomes in Arabidopsis thaliana TAIR10 assembly
1
2
3
Arabidopsis thaliana TAIR10 chr5:12322141-12322212 (72bp)
gene
mRNA
miRNA
Lib-1
Lib-2


BWA
mapping
(sequence db)
PMRD
mirBase
peak
detection
R

Expression

reads

K19

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**Probe**	**Chromosome**	**Start**	**End**	**Strand**	**No value**	**Feature**	**Description**	**Type**	**Orientation**	**Distance**	**LIB-1**	**LIB-2**	**LIB-3**	**LIB-4**
2	Chr1:21-87	1	21	87		NaN	null			Not found	0	80	104	50	76
3	Chr1:773-817	1	773	817		NaN	null			Not found	0	231	275	303	413
4	Chr1:18325-18349	1	18325	18349		NaN	null			Not found	0	26	27	13	24
5	Chr1:27895-27915	1	27895	27915		NaN	DCL1	dicer-like 1.[S	gene	overlapping	0	40	54	48	42
6	Chr1:44714-44746	1	44714	44746		NaN	AT1G01073	unknown prot	gene	overlapping	0	210	162	159	239
7	Chr1:50531-50625	1	50531	50625		NaN	AT1G01100	60S acidic rib	gene	overlapping	0	21	45	41	52
8	Chr1:55694-55798	1	55694	55798		NaN	AT1G01115	unknown prot	gene	downstream	826	130	120	160	274
9	Chr1:56431-56543	1	56431	56543		NaN	AT1G01115	unknown prot	gene	downstream	81	54	66	28	84
10	Chr1:77220-77494	1	77220	77494		NaN	AT1G01180	S-adenosyl-L	gene	upstream	462	140	125	218	269

profile

features

detection

Seqmonk
Bedtools

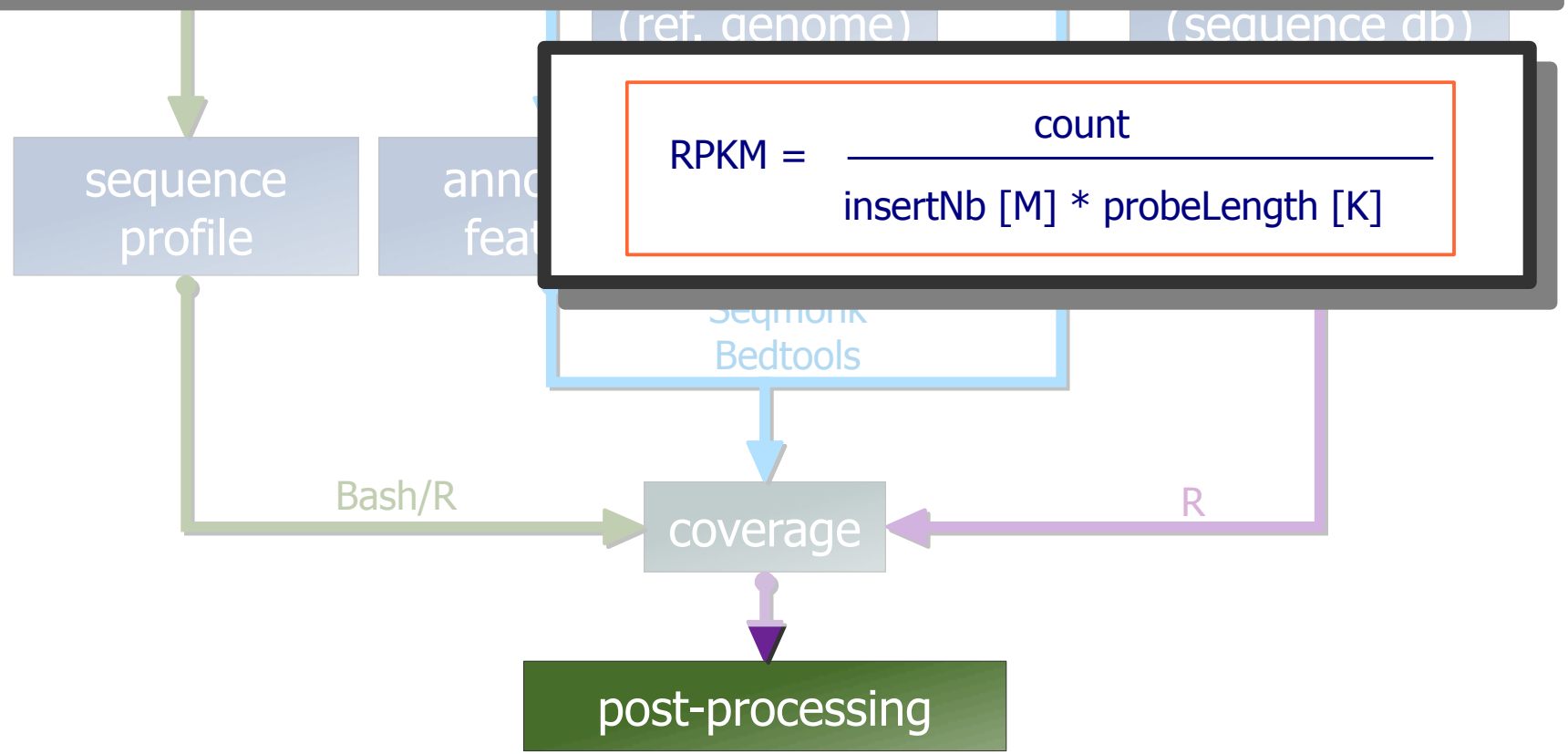
Bash/R

coverage

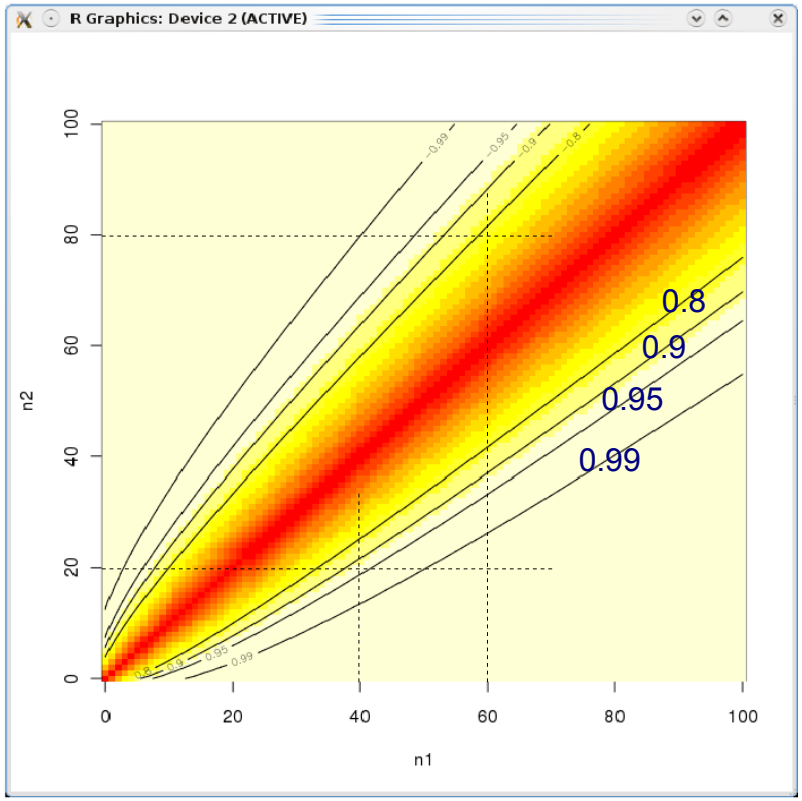
R

post-processing

RPKM LIB-1	RPKM LIB-2	RPKM LIB-3	RPKM LIB-4	score LIB-1 vs LIB-2	score LIB-1 vs LIB-3	score LIB-1 vs LIB-4	score LIB-2 vs LIB-3	score LIB-2 vs LIB-4	score LIB-3 vs LIB-4	rank
55.9	56.43	42.02	37.61	-1.27	0.28	0.08	0.24	0.06	0.79	11988
48.69	39.26	29.28	31.83	0.68	0.32	0.31	0.65	0.72	-1.13	12658
89.17	93.47	128.71	66.32	-1.16	-0.24	0.37	-0.27	0.25	0.03	3818
297.91	178.45	271.32	240.15	0	0.62	0.11	-0.01	-0.04	0.43	7122
10.35	17.22	24.3	18.15	-0.18	-0.02	-0.12	-0.27	-1.12	0.34	4344



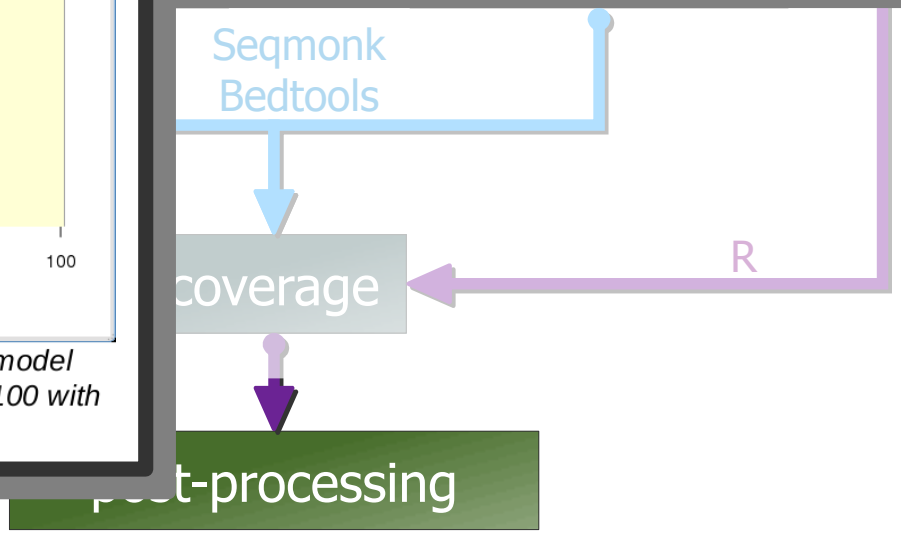
RPKM LIB-1	RPKM LIB-2	RPKM LIB-3	RPKM LIB-4	score LIB-1 vs LIB-2	score LIB-1 vs LIB-3	score LIB-1 vs LIB-4	score LIB-2 vs LIB-3	score LIB-2 vs LIB-4	score LIB-3 vs LIB-4	rank
55.9	56.43	42.02	37.61	-1.27	0.28	0.08	0.24	0.06	0.79	11988
48.69	39.26	29.28	31.83	0.68	0.32	0.31	0.65	0.72	-1.13	12658



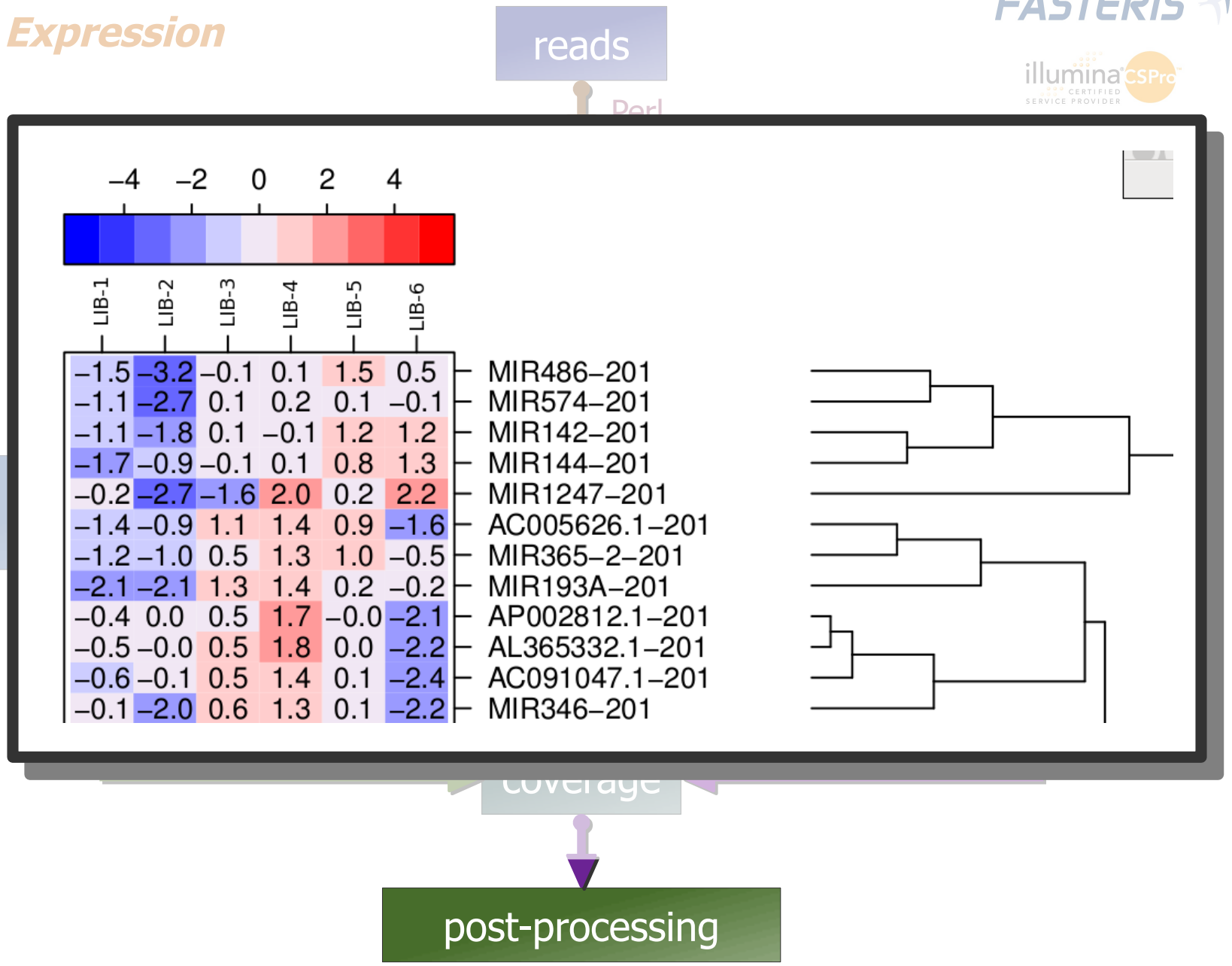
Overview of the scores obtained with the binomial model when comparing 2 counts (n1, n2) between 0 and 100 with (N1,N2) fixed to 1'000'000.

Comparison scores between pairs of libraries.
n1,n2~binomial distribution with same probability of event (p=(n1/N2+n2/N2)/2);

score~p(observing a count <n1 or >n2)



Expression


reads
Perl
−4 −2 0 2 4
LIB-1 LIB-2 LIB-3 LIB-4 LIB-5 LIB-6
−1.5 −3.2 −0.1 0.1 1.5 0.5 MIR486−201
−1.1 −2.7 0.1 0.2 0.1 −0.1 MIR574−201
−1.1 −1.8 0.1 −0.1 1.2 1.2 MIR142−201
−1.7 −0.9 −0.1 0.1 0.8 1.3 MIR144−201
−0.2 −2.7 −1.6 2.0 0.2 2.2 MIR1247−201
−1.4 −0.9 1.1 1.4 0.9 −1.6 AC005626.1−201
−1.2 −1.0 0.5 1.3 1.0 −0.5 MIR365−2−201
−2.1 −2.1 1.3 1.4 0.2 −0.2 MIR193A−201
−0.4 0.0 0.5 1.7 −0.0 −2.1 AP002812.1−201
−0.5 −0.0 0.5 1.8 0.0 −2.2 AL365332.1−201
−0.6 −0.1 0.5 1.4 0.1 −2.4 AC091047.1−201
−0.1 −2.0 0.6 1.3 0.1 −2.2 MIR346−201
coverage
post-processing

Virus identification

Sequencing by siRNA: a novel generic tool for virus discovery

Kreuze et al. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. ***Virology*** **388: 1-7**

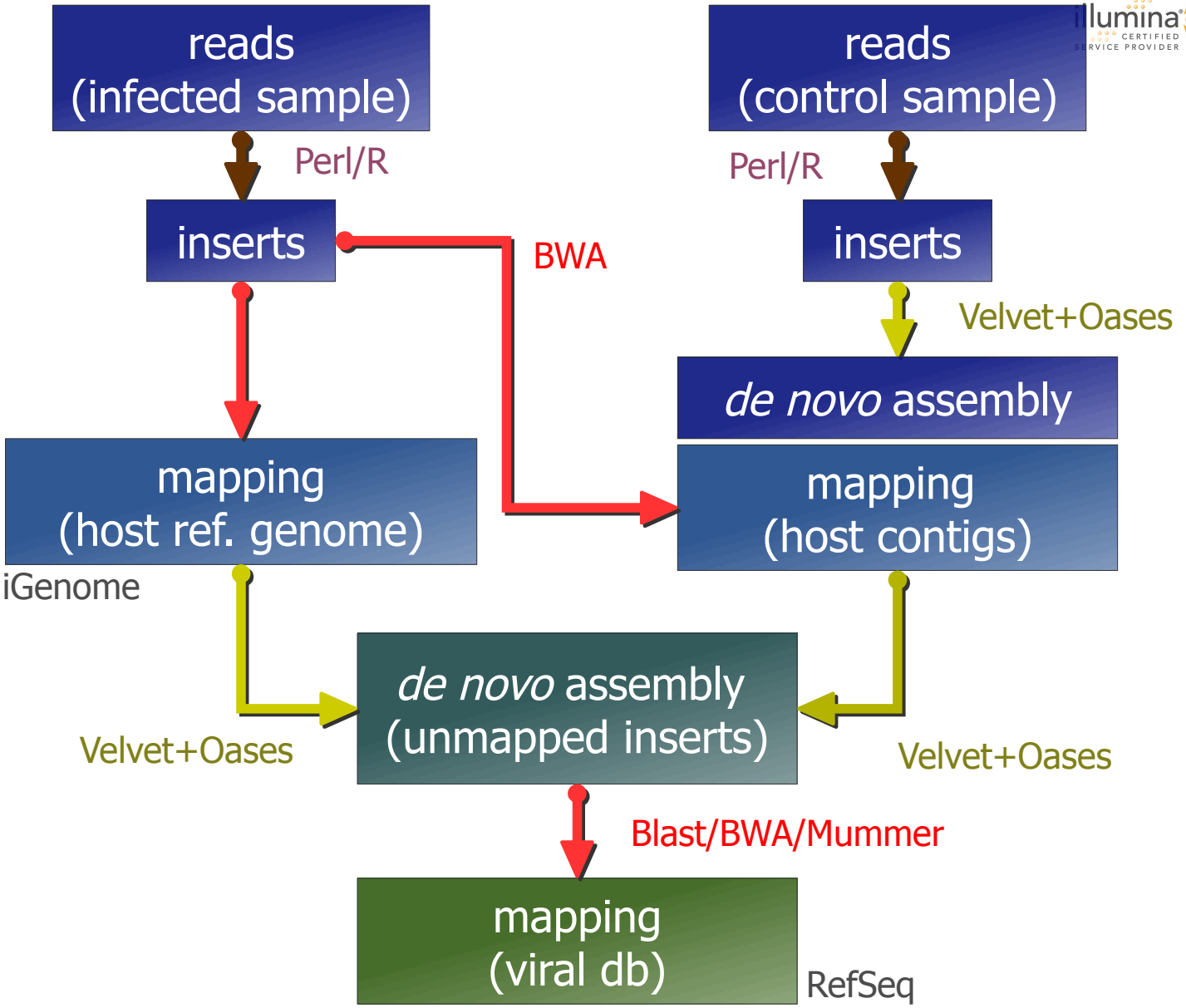
SiRNAs:

- class of dsRNAs of 20-25 nts
- involved in post-transcriptional gene silencing
- endogenous or exogenous

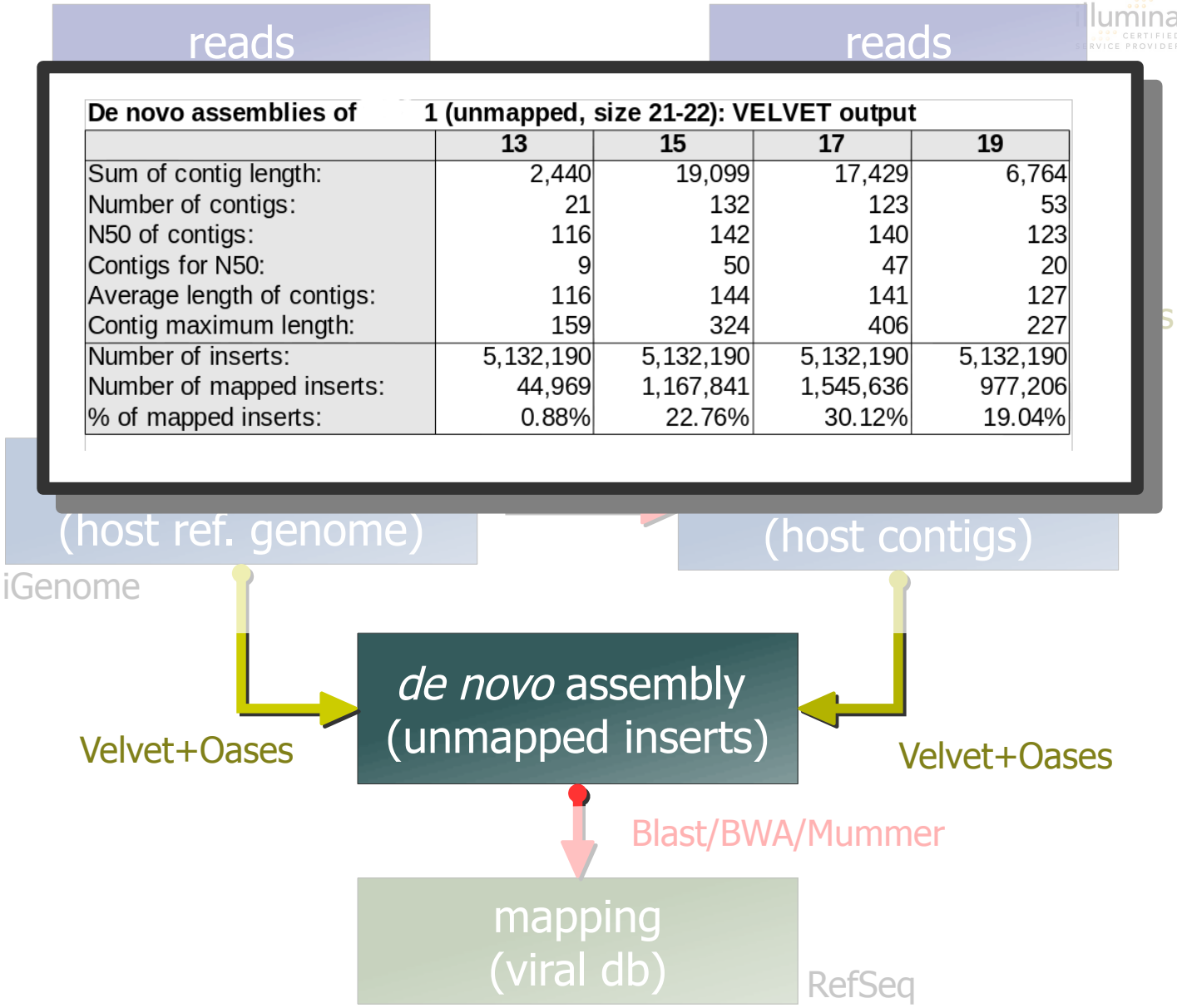
→ synthetic dsRNA introduced into cells can induce silencing of specific genes of interest

→ viral infection: presence of viral dsRNA leading to siRNAs that participate in the cell antiviral response;

Virus assembly pipelines



Virus assembly



De novo assemblies of	1 (unmapped, size 21-22): VELVET output			
	13	15	17	19
Sum of contig length:	2,440	19,099	17,429	6,764
Number of contigs:	21	132	123	53
N50 of contigs:	116	142	140	123
Contigs for N50:	9	50	47	20
Average length of contigs:	116	144	141	127
Contig maximum length:	159	324	406	227
Number of inserts:	5,132,190	5,132,190	5,132,190	5,132,190
Number of mapped inserts:	44,969	1,167,841	1,545,636	977,206
% of mapped inserts:	0.88%	22.76%	30.12%	19.04%

Virus assembly

reads

reads

Aligned bases from the virus	% of aligned bases from the virus	Aligned bases in the set of contigs	% of aligned bases in the set of contigs	Size of the virus	GeneBank AC	Definition
14626	75.97%	24311	47.65%	19252	DQ151548	Citrus tristeza virus strain T318A, complete genome.
14197	73.55%	23217	45.50%	19302	AB046398	Citrus tristeza virus genomic RNA, complete genome, seedling
14191	73.63%	22653	44.40%	19273	FJ525435	Citrus tristeza virus isolate NZRB-M17, complete genome.
13938	72.33%	21802	42.73%	19270	FJ525434	Citrus tristeza virus isolate NZRB-TH30, complete genome.
13933	72.40%	22156	43.42%	19245	GQ454869	Citrus tristeza virus strain HA18-9, complete genome.
13701	71.18%	22107	43.33%	19249	AF001623	Citrus tristeza virus, complete genome.
12889	66.95%	22505	44.11%	19251	EU937519	Citrus tristeza virus strain VT, complete genome.
12868	66.84%	22718	44.52%	19253	HM573451	Citrus tristeza virus isolate Kpg 3, complete genome.
12792	66.43%	19923	39.05%	19255	FJ525433	Citrus tristeza virus isolate NZRB-TH28, complete genome.
12688	65.99%	21801	42.73%	19226	U56902	Citrus tristeza virus p346, 54-kDa RNA dependent RNA polyme

Velvet+Oases

de novo assembly (unmapped inserts)

Velvet+Oases

Blast/BWA/Mummer

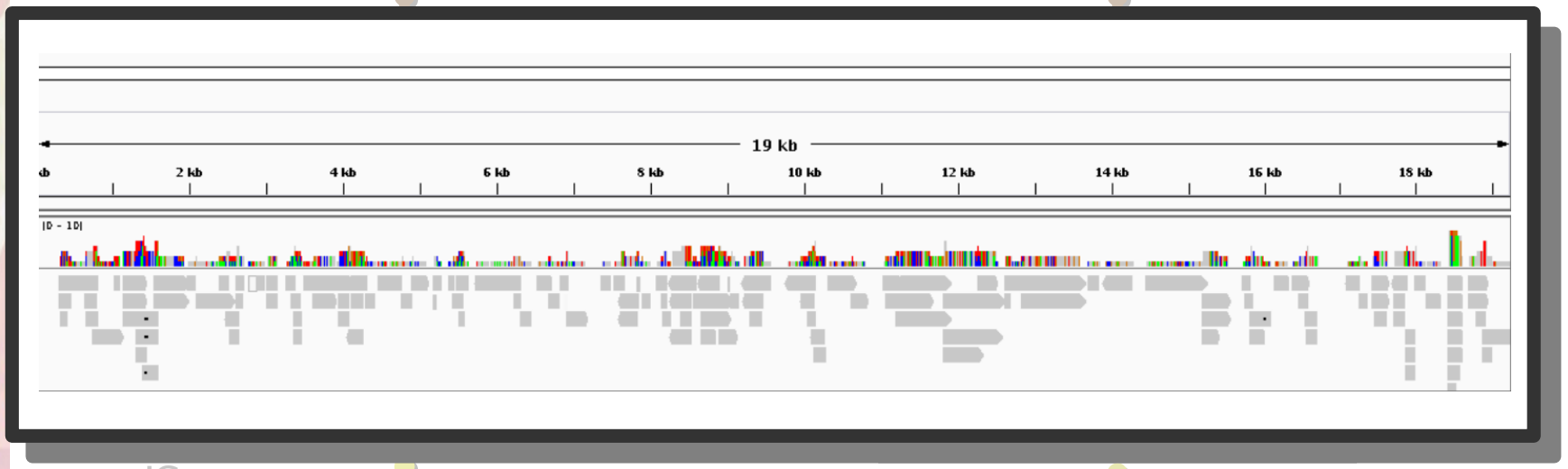
mapping (viral db)

RefSeq

Virus assembly

reads (infected library)

reads (control library)



iGenome

Velvet+Oases

de novo assembly (unmapped inserts)

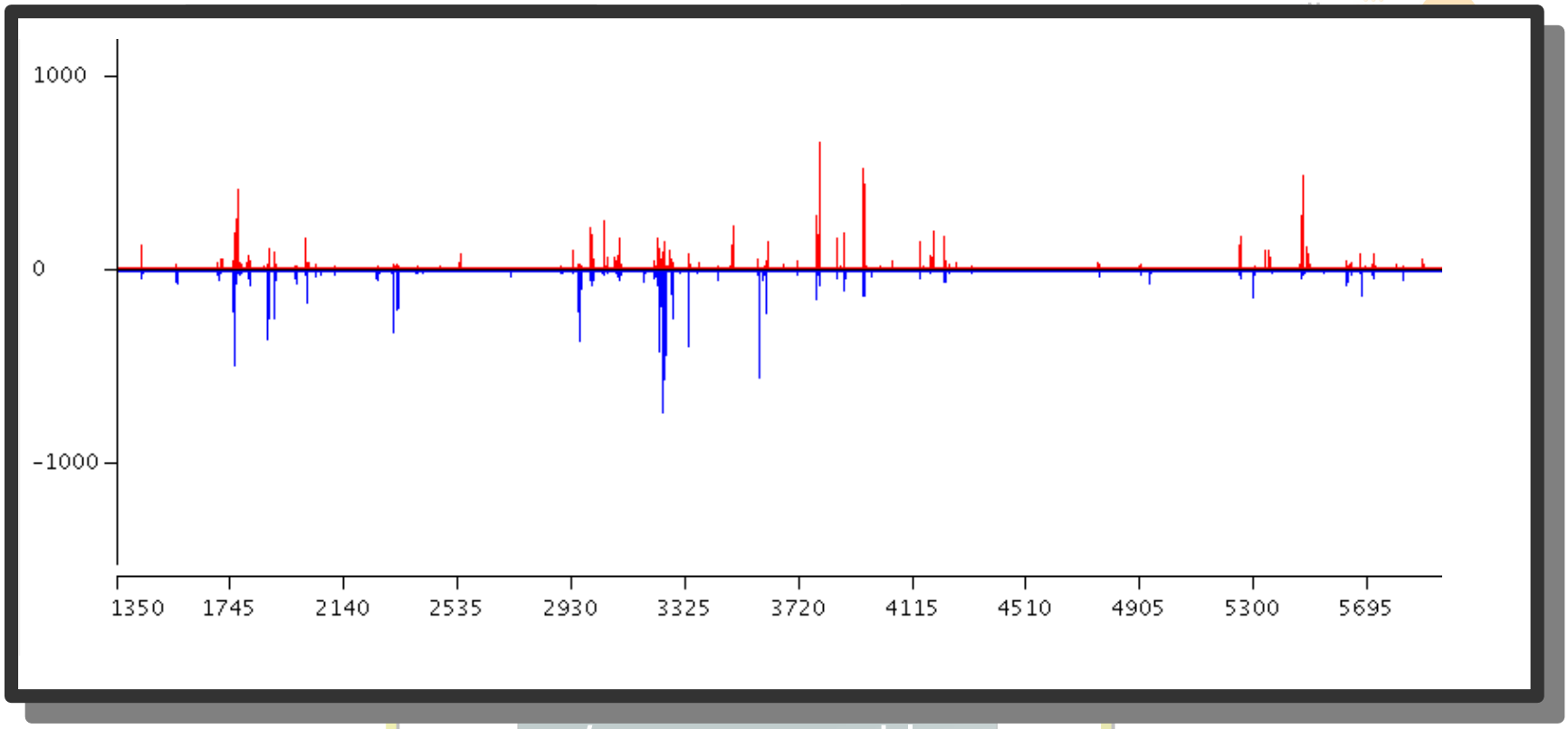
Velvet+Oases

Blast/BWA/Mummer

mapping (viral db)

RefSeq

Virus assembly



de novo assembly (unmapped inserts)

Velvet+Oases

Velvet+Oases

Blast/BWA/Mummer

mapping (viral db)

RefSeq

FASTERIS

illumina CSPro CERTIFIED SERVICE PROVIDER

Thank you for your attention