RAP: RNA-Seq Analysis Pipeline

A cloud-based NGS web application

RNASeq2015

Mattia D'Antonio

CINECA Roma

m.dantonio@cineca.it

ELIXIR-ITA

Rome, 22th October, 2015

What is RAP?

RNA-Seq Analysis Pipeline

- A fully automated RNA-Seq pipeline
  - To study gene expression in a wide variety of conditions
  - Easy access to computational clusters
  - Comprehensive and complex analysis pipeline
  - Integrated analysis tools

- A user friendly WEB service
  - Easy access to bioinformatics tools
  - No technical requirements
  - A better and powerful results visualization

Why use RAP?

A simple web interface allows the user to:

- Create, customize and monitor an analysis
- Browse, filter and download the results
- Avoid any software installation and configuration
- Updated bioinformatics software
- Access to computational resources in cloud

RNA-Seq data size

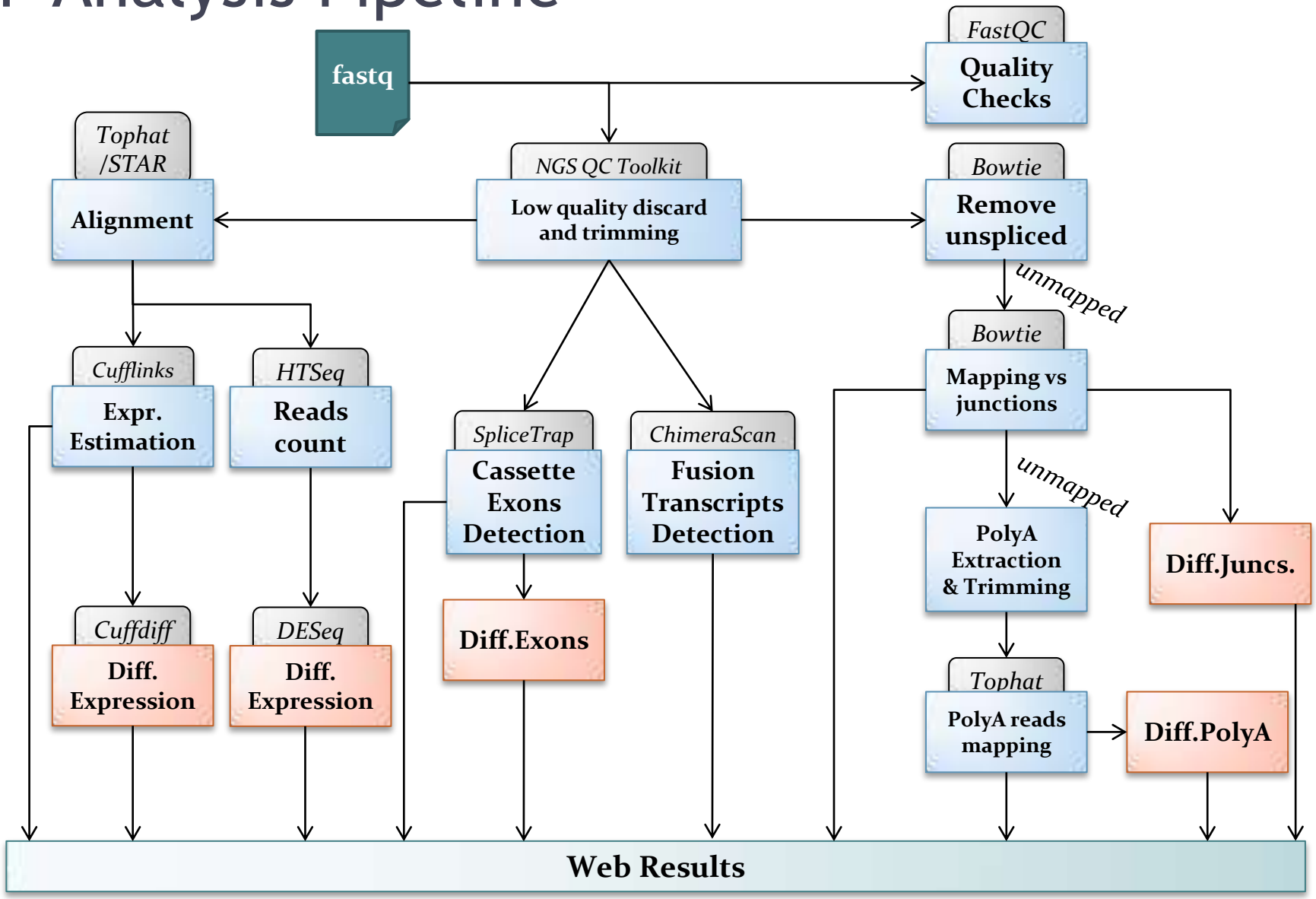
NGS applications are resource-hungry

- RNA-Seq in particular

- HiSeq 2000 Output:
  - 300 Gb (fastq)
  - 375 Million/lane PE reads
- Increased size due to replicates and PE

- This huge quantity of data requires computational clusters
  - Support of Computational Centers (e.g. CINECA)

RAP Analysis Pipeline



Workflow modules

- Preliminary modules
  - Quality checks (*FastQC*)
  - Reads filtering (*NGS QC Toolkit*)
- Main branch
  - Alignment (*Tophat/STAR*)
  - Transcripts reconstruction and quantification (*Cufflinks / HTSeq*)
  - Differential expression (*Cuffdiff / DESeq*)
- Splicing Junctions and PolyA
  - Junctions mapping (*Bowtie / Tophat*)
  - PolyA tags (*Bowtie*)
- Cassette Exons (*SpliceTrap*)
- Fusion Transcripts (*ChimeraScan*)
- Result
  - Results visualization
  - Pathway analysis, Plots

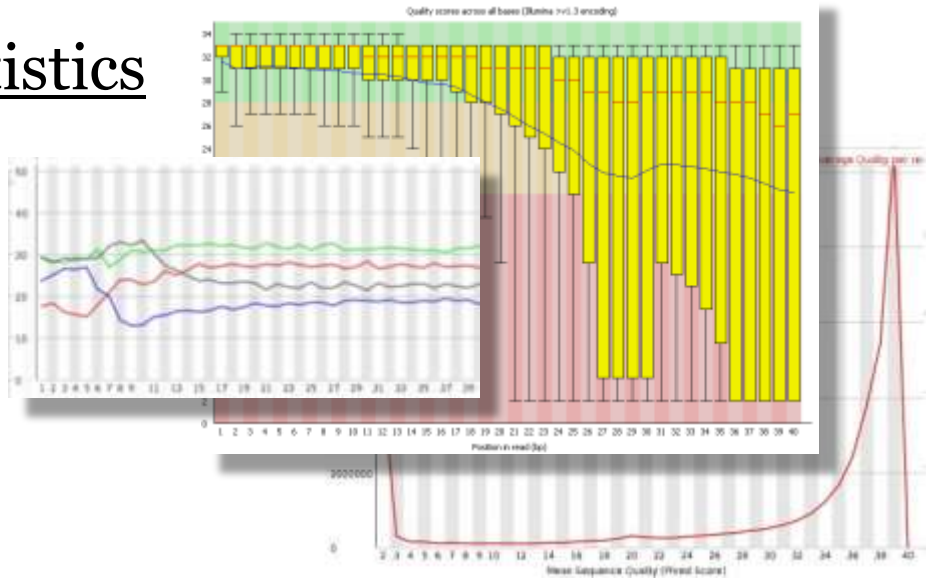
The RAP Workflow

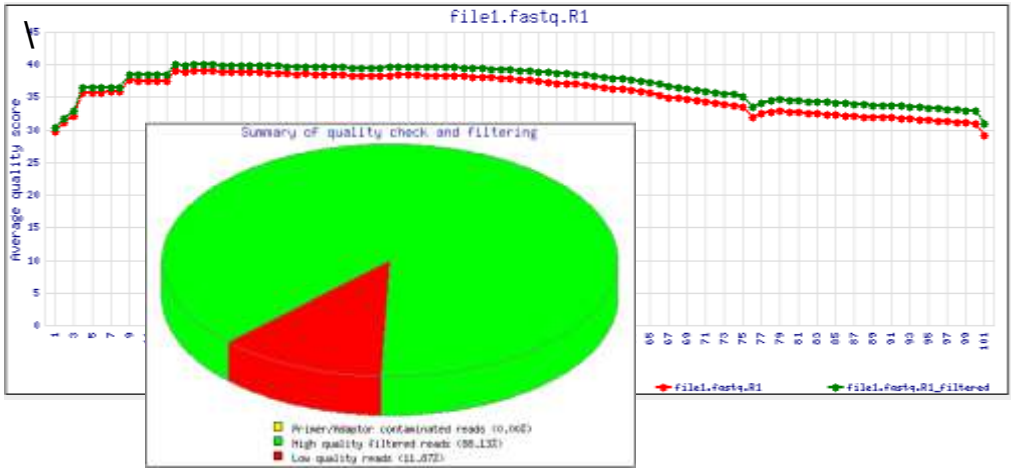
Chapter 1

Quality assessment

FastQC gives <u>quality statistics</u>

- base sequence quality
- base sequence content,
- sequence GC content



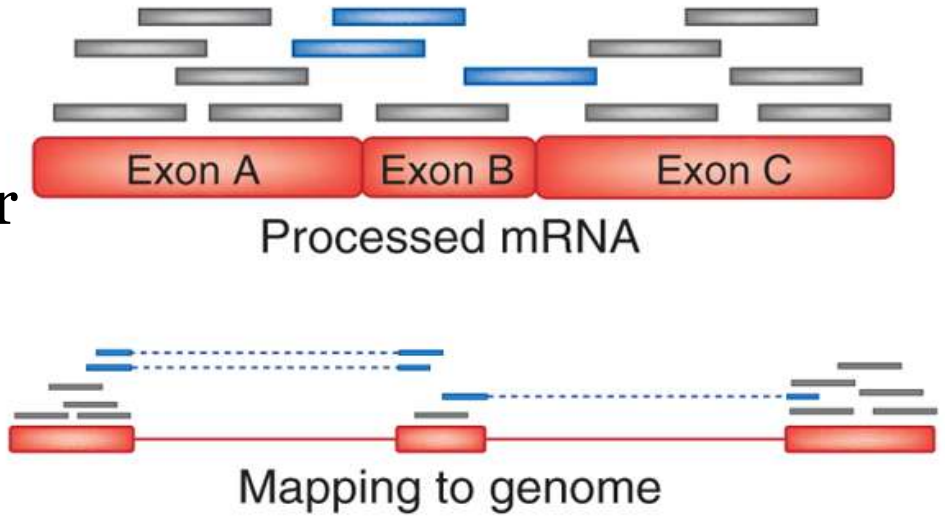


NGS QC Toolkit gives the possibility to <u>trim and discard</u> low quality bases and sequences

Reads alignment - Tophat

- High quality Reads are aligned to the genome by **TopHat**
- Tophat identifies exon-exon splice junctions
  - uses the mapping program Bowtie
- RNA-Seq reads can contiguously align to the genome
  - unspliced reads

TopHat is a spliced aligner
  - input reads are split into smaller segments mapped independently



Reads alignment - STAR

- As alternative, reads can be aligned to the genome by using **STAR**
- Ultrafast universal RNA-seq aligner
  - Outperforms other aligners by a factor of > 50
  - 550 million 2 x 76bp PE reads per hours on 12-core
  - Very good scalability
  - Huge RAM requirements
  - Good accuracy (maybe better than Tophat? However not so worst)
- Splicing detection requires a bit of work (to be done)
  - Indexes building should consider read length

Transcript assembly and quantification

Results of alignment are given to **Cufflinks** to:

- Reconstruction of a parsimonious set of transcripts
- Estimation of the relative abundances of transcripts
  - Expression levels are measured in FPKM
    - Fragments Per Kilobase of exon per Million mapped fragments

FPKM = fragments count with the application of length-based normalization

Gene raw counts

Results of alignment are also analyzed by **HTSeq**

- Raw reads count at gene level
- Useful for differential expression modules like DESeq
- Raw counts are in difficult with
  - multi-mapping reads
  - ambigous alignment

Splicing junctions

To determine expressed junctions:

1. Alignment to the reference genome with **Bowtie**
   - unspliced reads (intra-exonic reads) are discarded
   - unmapped reads are potentially spliced

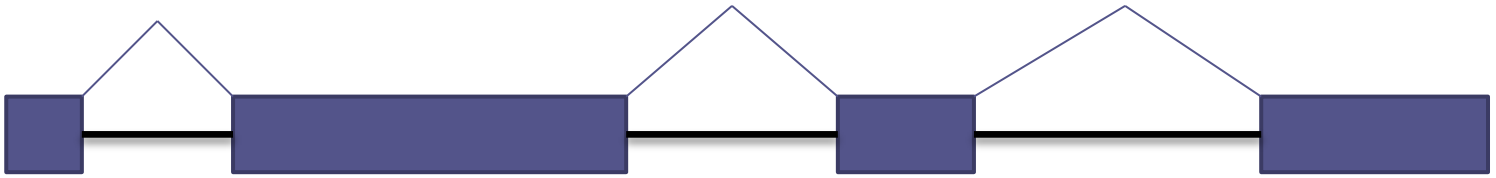
2. Alignment to a custom built splice junctions library
   - obtained from RefSeq

RefSeq = database of annotated and curated nucleotide sequences

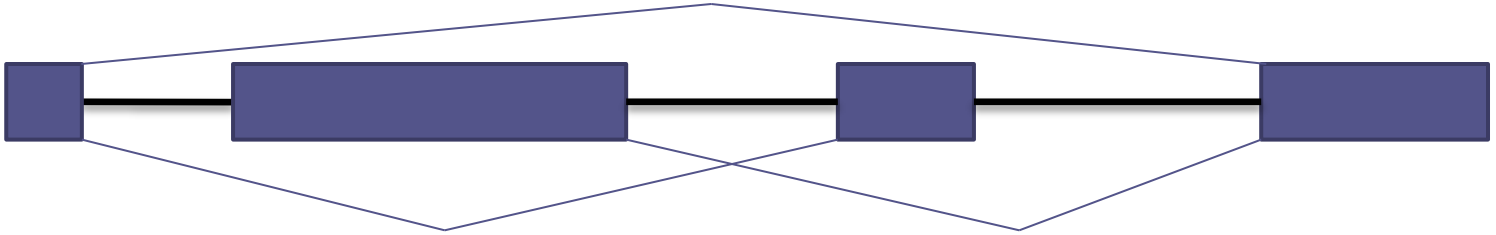
Splice Junctions library

This library contains:

- known splice junctions
  - Each pair of consecutive exon boundaries
  - 200.000 junctions from RefSeq



- potential splice junctions
  - Through a combinatorial exon skipping procedure
  - ~ 2 million junctions

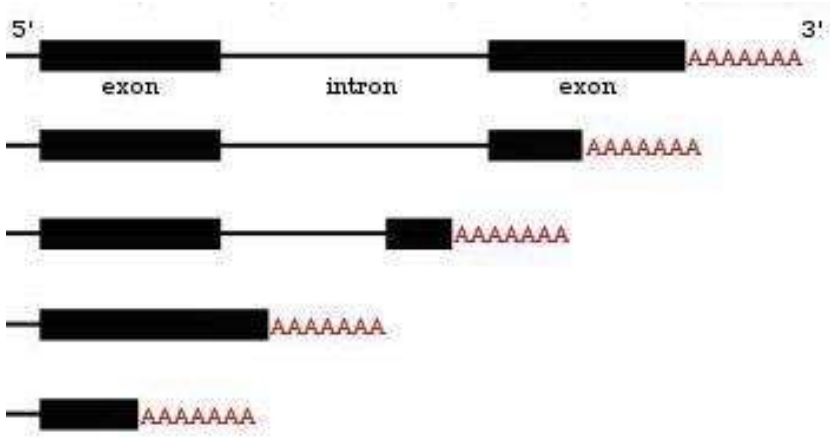


Polyadenylation sites

A polyA tag cannot be mapped to the genome and polyA tail should be removed before mapping

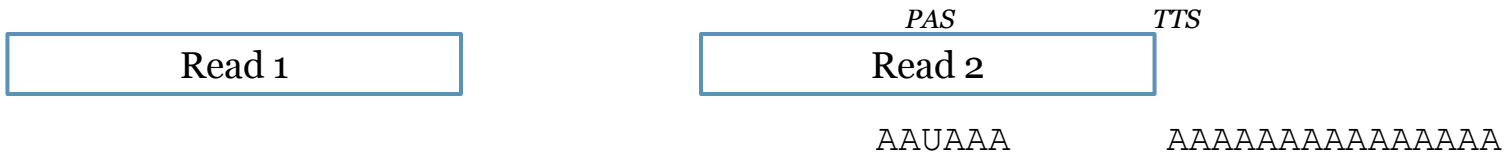
```
TGACTGACTGATACTGACACACTGATCGATCG
||||||||||||||||||||
ACTGACTGTCTATGACTGAAAAAAAAAAAAAA
```

PolyA tags mark transcription termination sites (TTS)



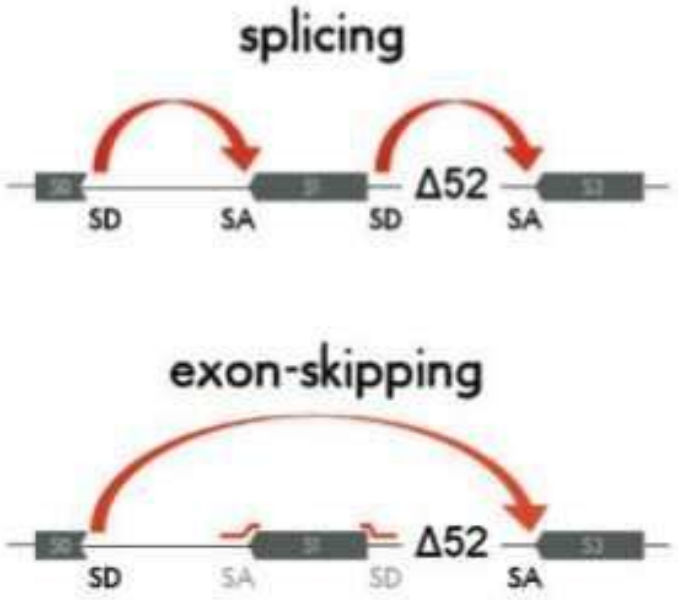
PolyA detection

1. Extraction of unmapped reads
   - From Tophat and/or junctions
2. Extraction of polyA according to the read strand
   - Consider directional
3. PolyAs tails are trimmed
   - Trimming position is annotated
4. Mapping to the genome with **TopHat**
5. PolyA sites annotation
6. Detection of Polyadenylation Signals (PAS)



Cassette Exons

An exon may be spliced out of the primary transcript or retained
(CE or exon skipping)



SpliceTrap is a statistic tool for

- cassette exons identification
- quantifying exon inclusion ratios

SpliceTrap

- ✓ Utilizes a comprehensive exon trio db
- ✓ Aligns reads using bowtie (or RMAP)
- ✓ Exploits the abundances and positions of the read
- ✓ Tests the hypothesis whether the middle exon is alternatively spliced or not
- ✓ Calculates inclusion ratios for cassette exons

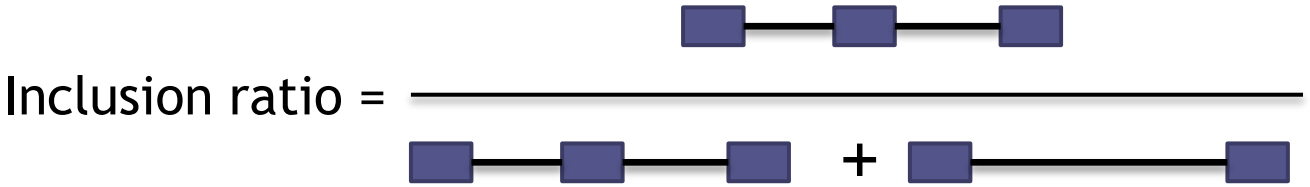
Inclusion ratios

The exon **inclusion ratio** is

the expression level of the **inclusion isoform**

divided by

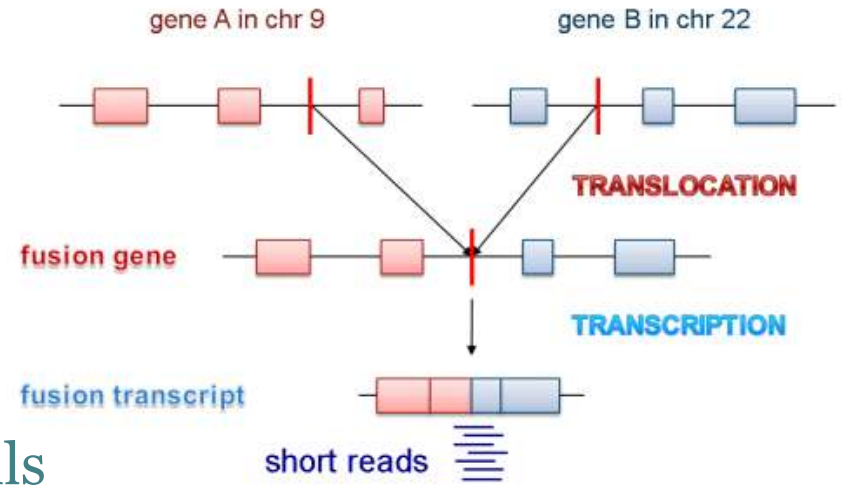
total expression level of both isoforms
(inclusion and skipped)



Chimeric Transcripts

Chimeric RNA is encoded by

- ✓ a fusion gene
- ✓ trans-splicing



- ❖ Often produced by cancer cells
- ❖ Chimeras can be functional!

ChimeraScan detects gene fusions in paired-end datasets by mapping to both genome and transcriptome to find **fusion breakpoints** candidates

ChimeraScan

Reads that map to different isoforms of the same gene are discarded

The others are fusion breakpoints candidates

Potential chimeras are filtered to remove artefacts

- ✓ Low coverage
- ✓ Low expressed chimeric transcripts

DE transcripts

Cuffdiff finds significant changes in

- transcripts expression
- Splicing events
- promoters

- Variance is modelled as a function of the mean fragment count across replicates

- It is known to use an overly conservative model
  - With few replicates or high variability may not get any significant calls

DE genes

Another option is **DESeq**

- DE at gene level

- Gene raw counts in input
  - e.g from HTSeq

- More robust with few biological replicates

- It divides variability in
  - Raw variance (from biological variability)
  - Shot noise (from counts uncertainty)

DE polyA, junctions, exons

Other differential analyses are performed to compare results obtained at several steps

- Alternative polyadenylation sites
  - Mapping counts on poly(A) sites
  - DESeq
- Alternative splicing junctions
  - Mapping counts on junctions
  - DESeq
- Alternative Exon inclusion ratios
  - Inclusion ratios
  - χ^2 test

Available organisms

Several organisms are available in RAP:

- Homo Sapiens — hg18, hg19
- Mus Musculus — mm9, mm10
- Rattus Norvegicus — rn4
- Drosophila Melanogaster — dm3
- Saccharomyces cerevisiae — sacCer3
- Zea mays — maize2, maize3, m017

The RAP Web Interface

Chapter 2

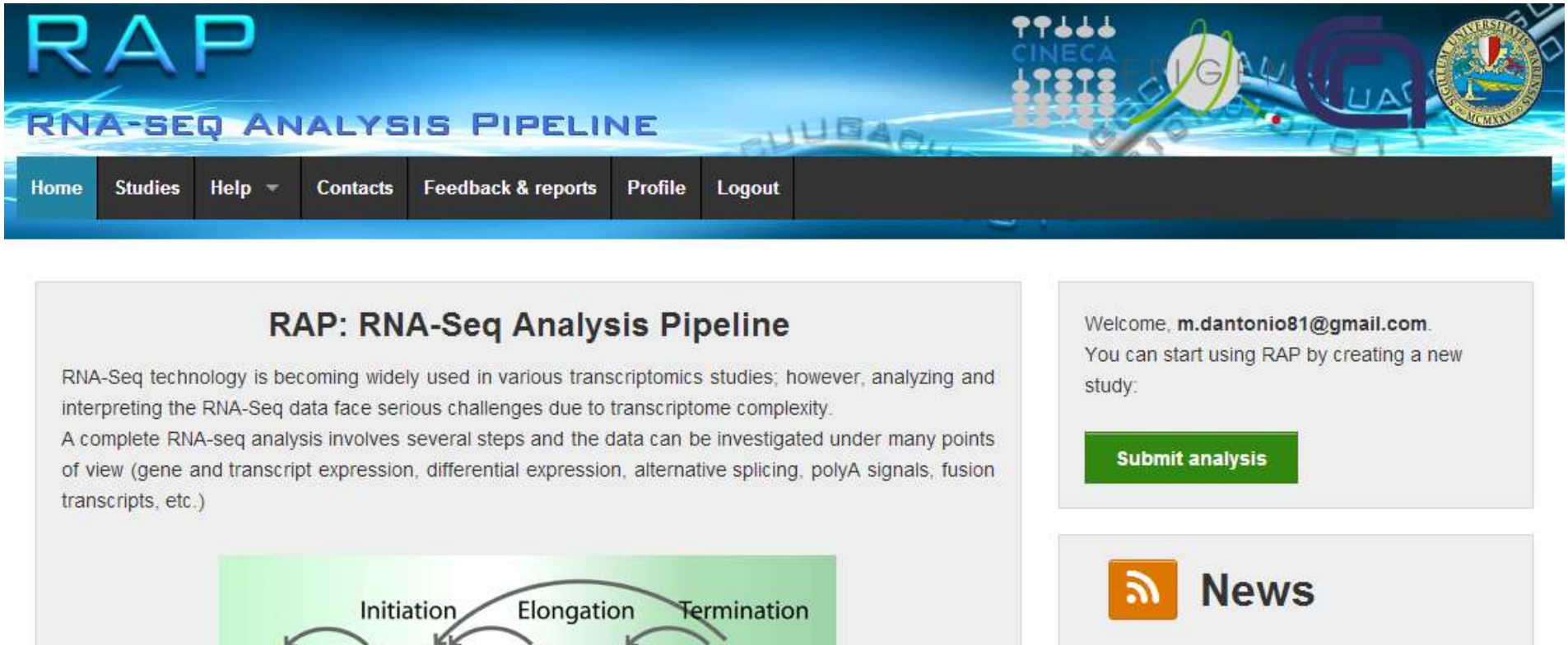
Test user accounts

- username: ***rnaseq2015-n@cineca.it***
  - rnaseq2015-1@cineca.it
  - rnaseq2015-2@cineca.it
  - …
  - rnaseq2015-25@cineca.it
- password: ***elixir-ita***

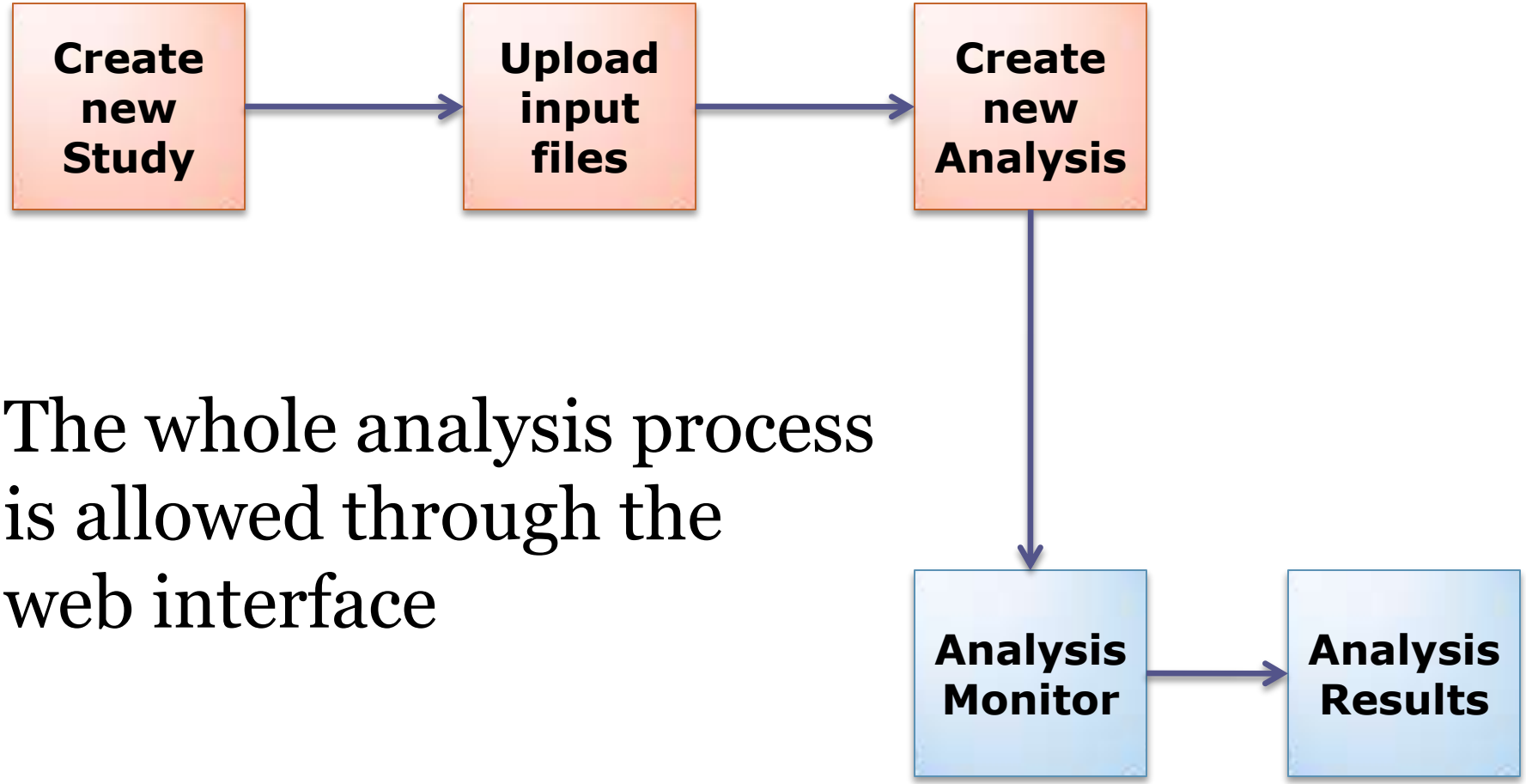
The web interface

A web 2.0 interface based on

HTML5, CSS3, jQuery, Foundation 4

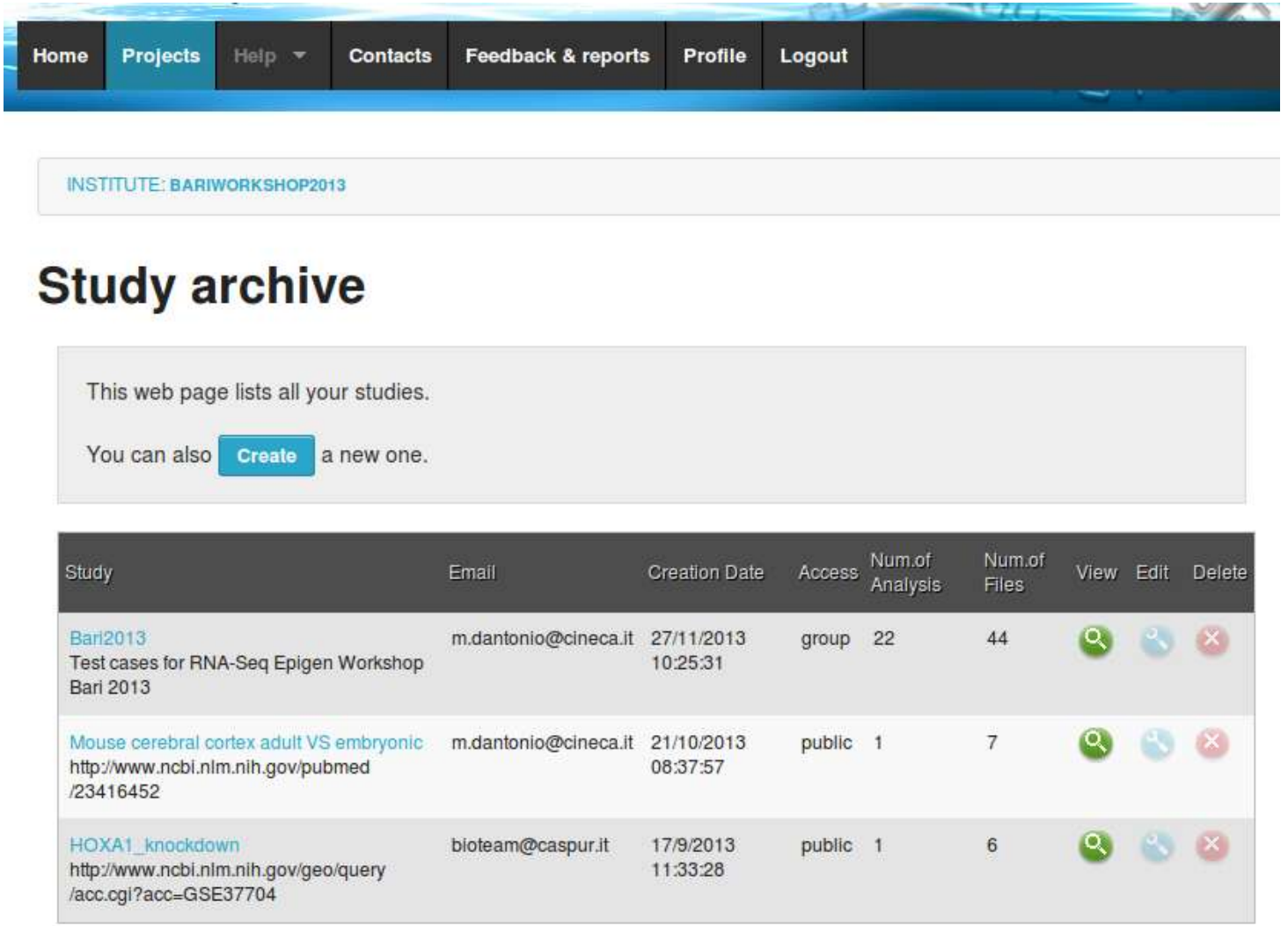


Data submission



The whole analysis process is allowed through the web interface

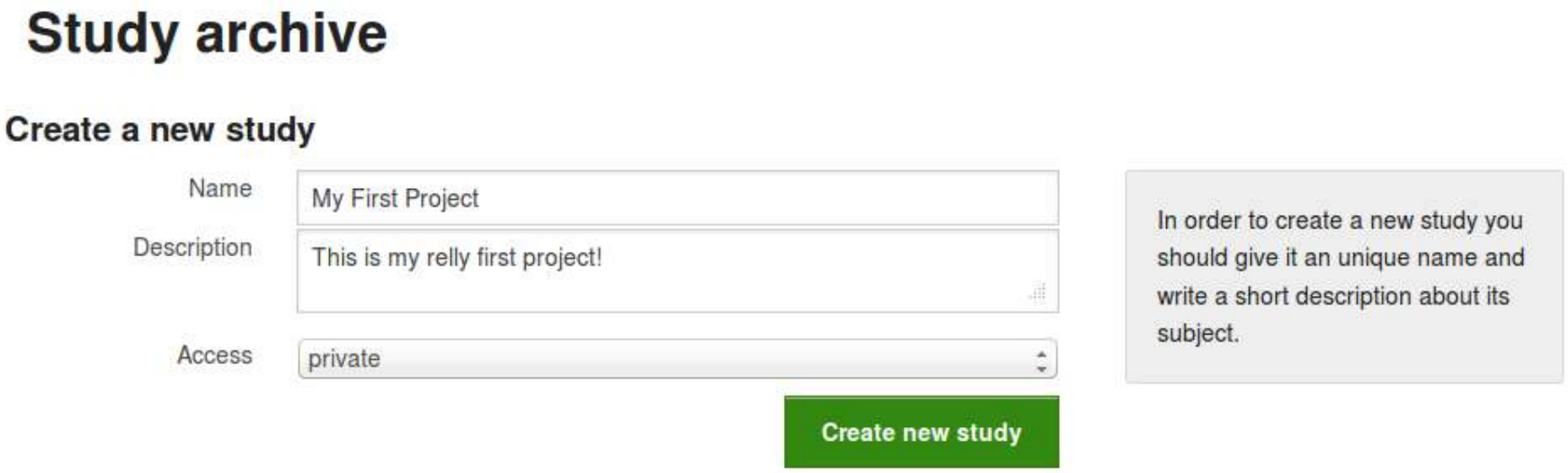
Projects list



Create a study (or Project)

To create a new study you should give it:

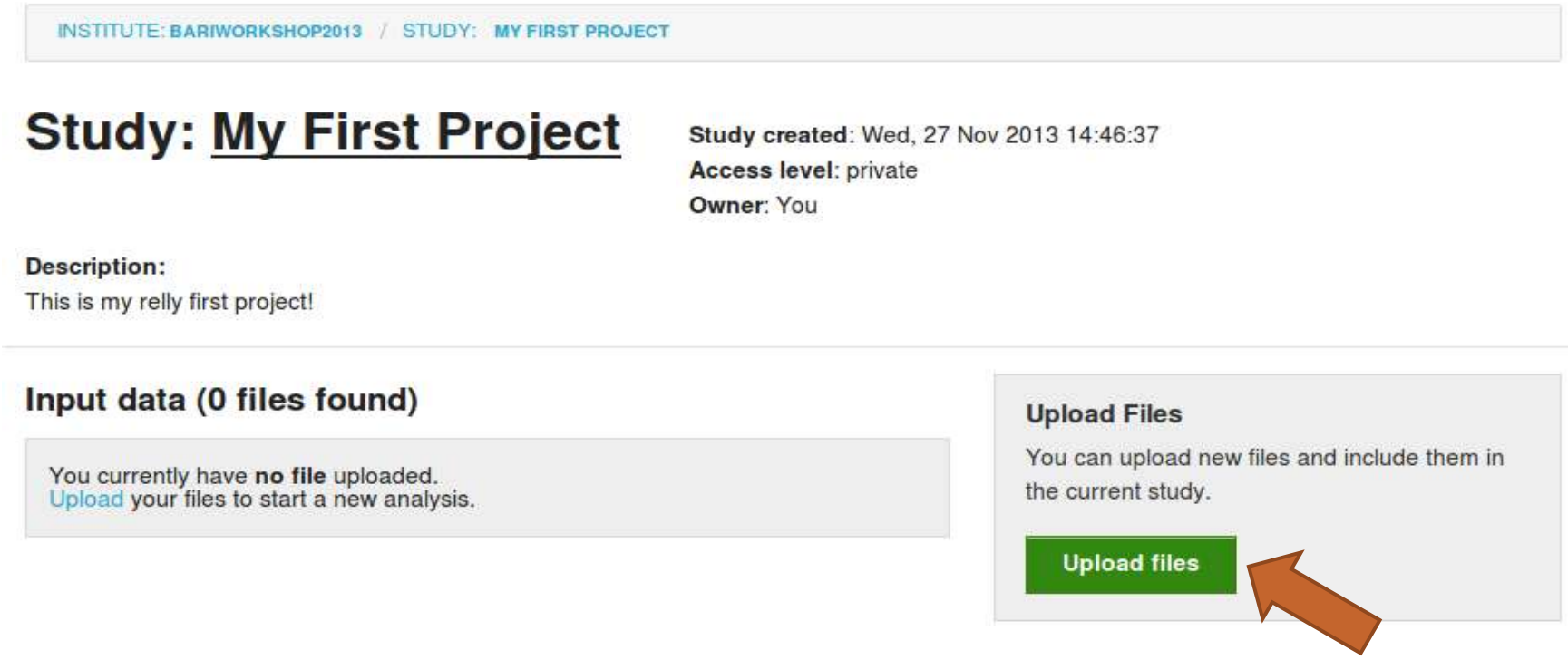
- ✓ a unique name
- ✓ a short description
- ✓ an access level



My first study

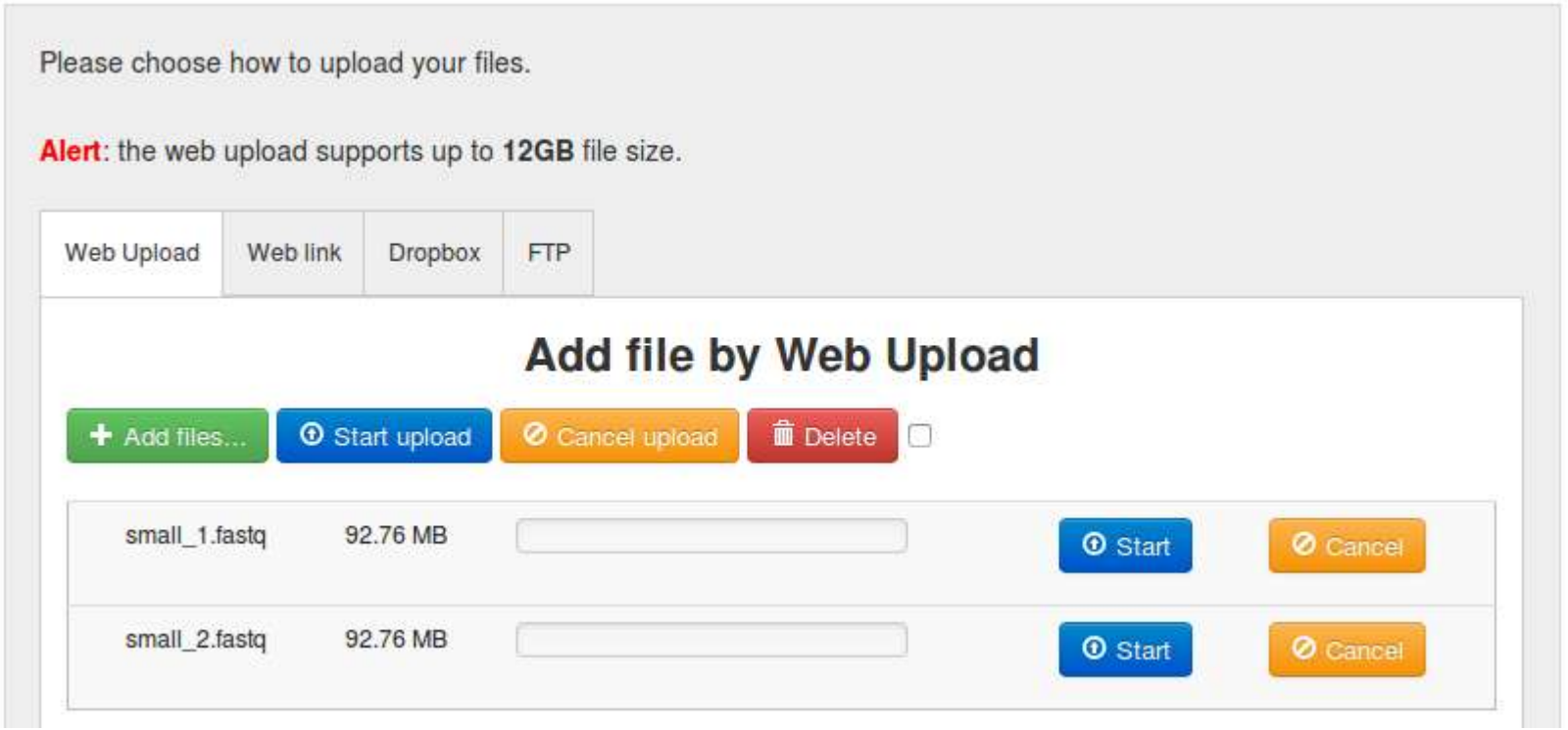
A new study is ust an empty container…

… you need input files



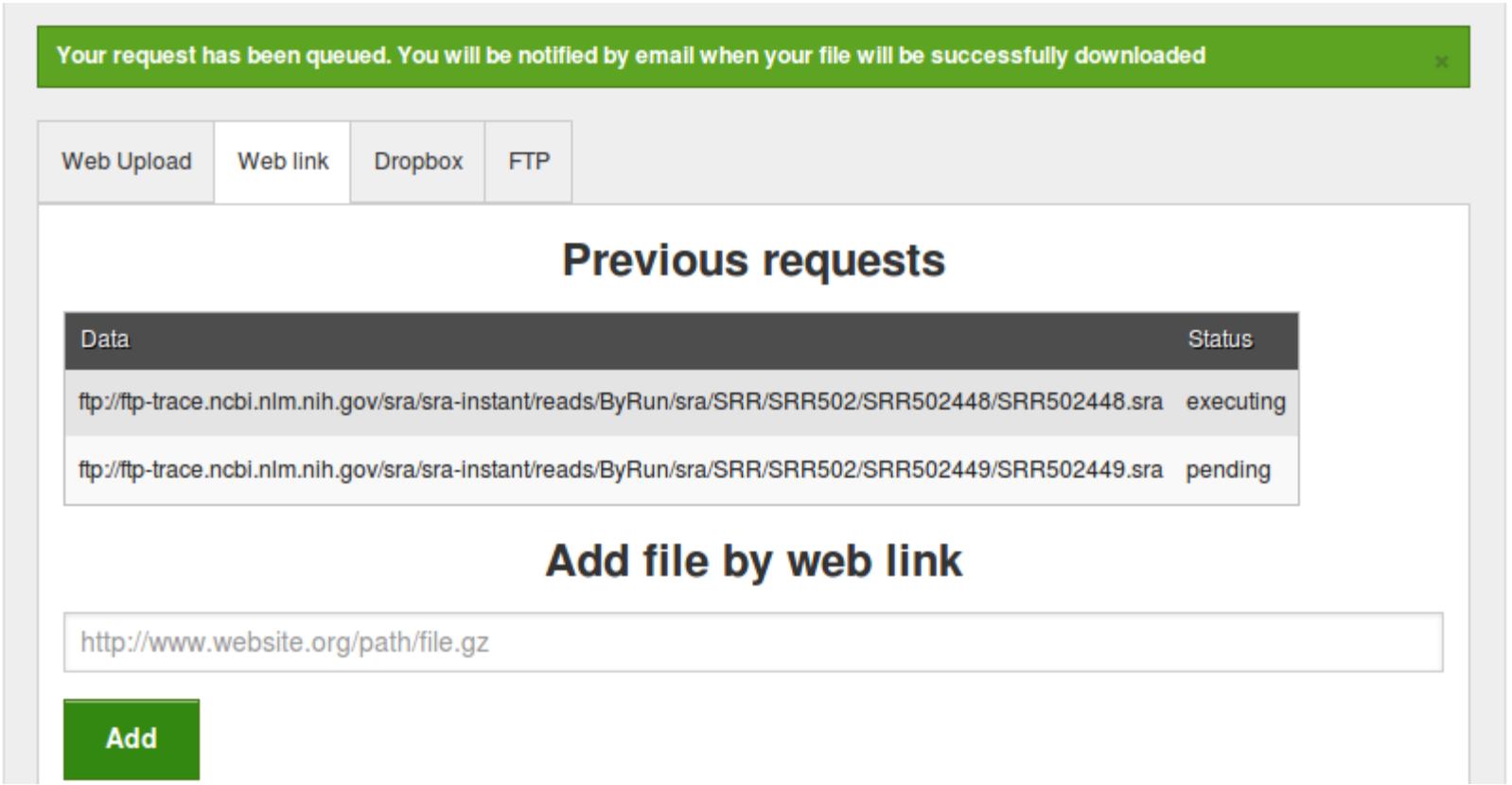
Upload input

Several upload options

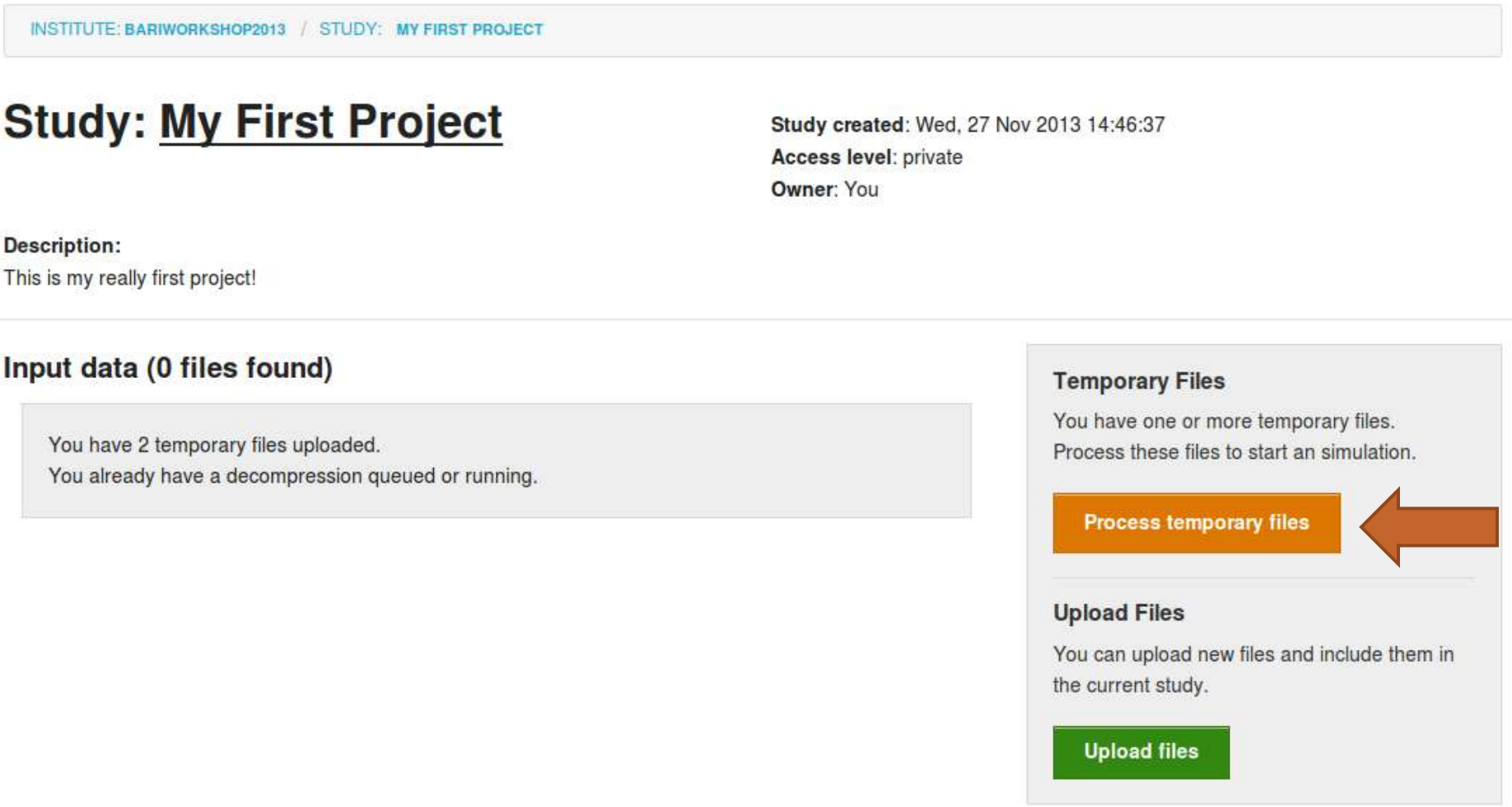


Web Upload

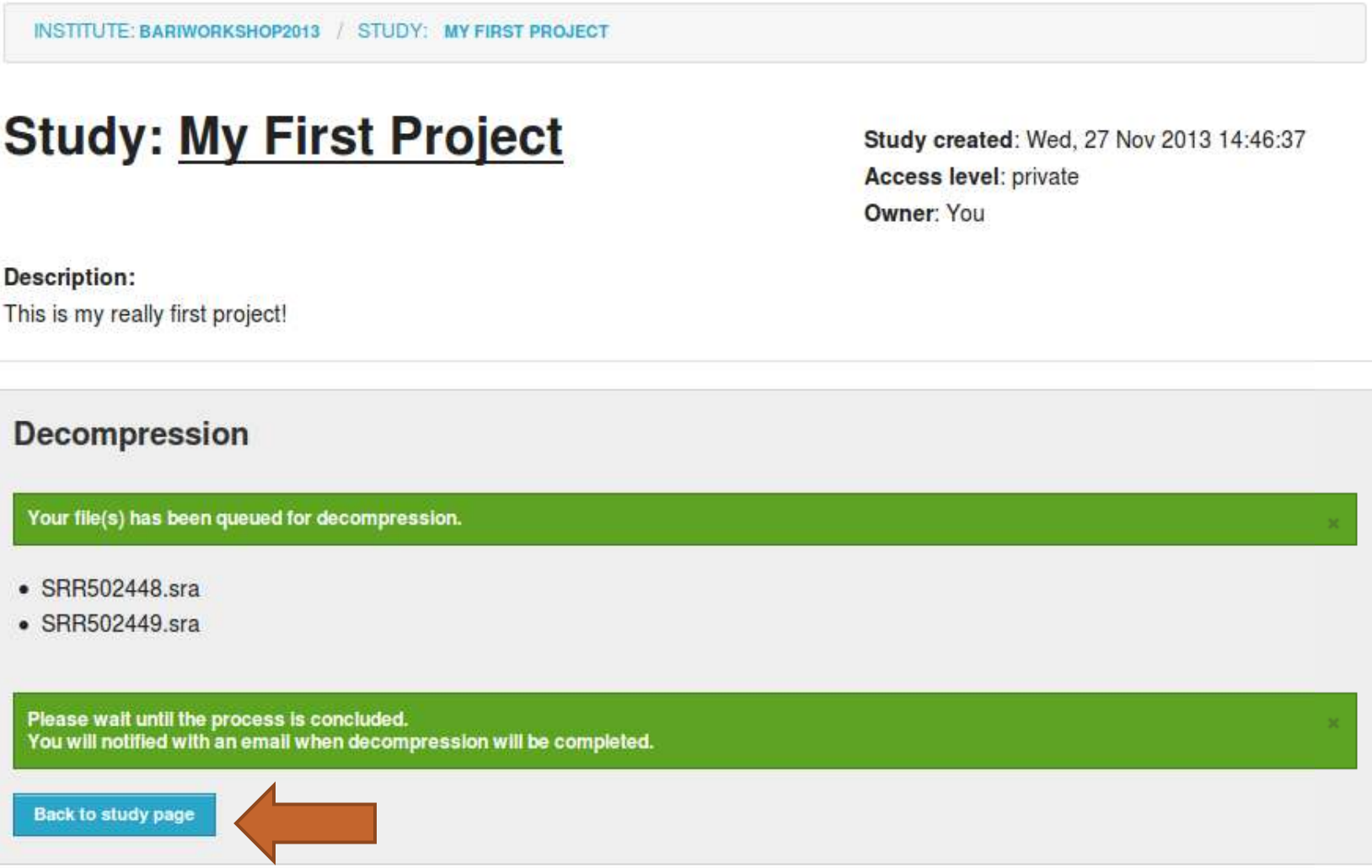
No limitation to file size, ansynchronous



Preprocessing - temporary files

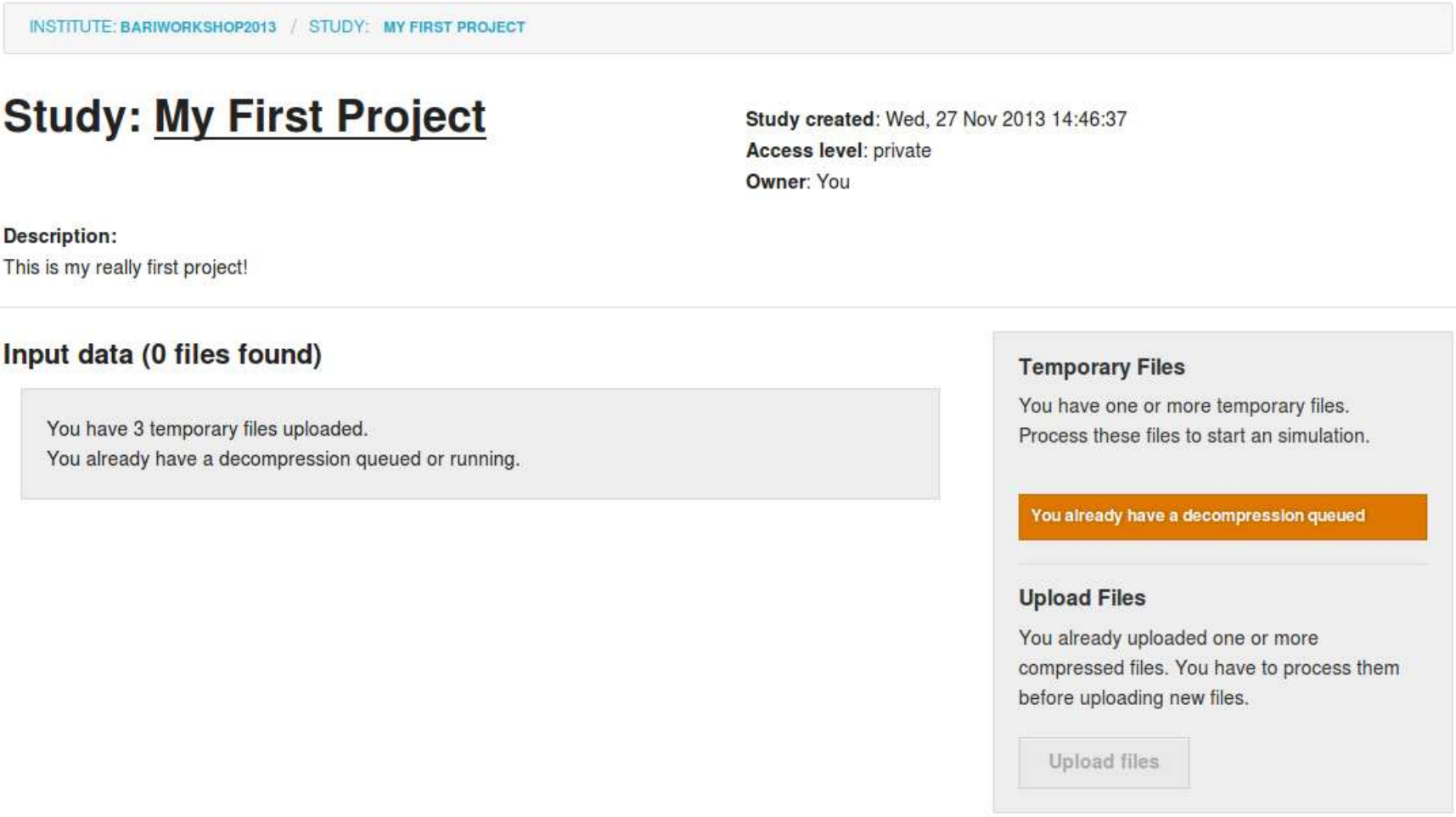


Preprocessing - decompression

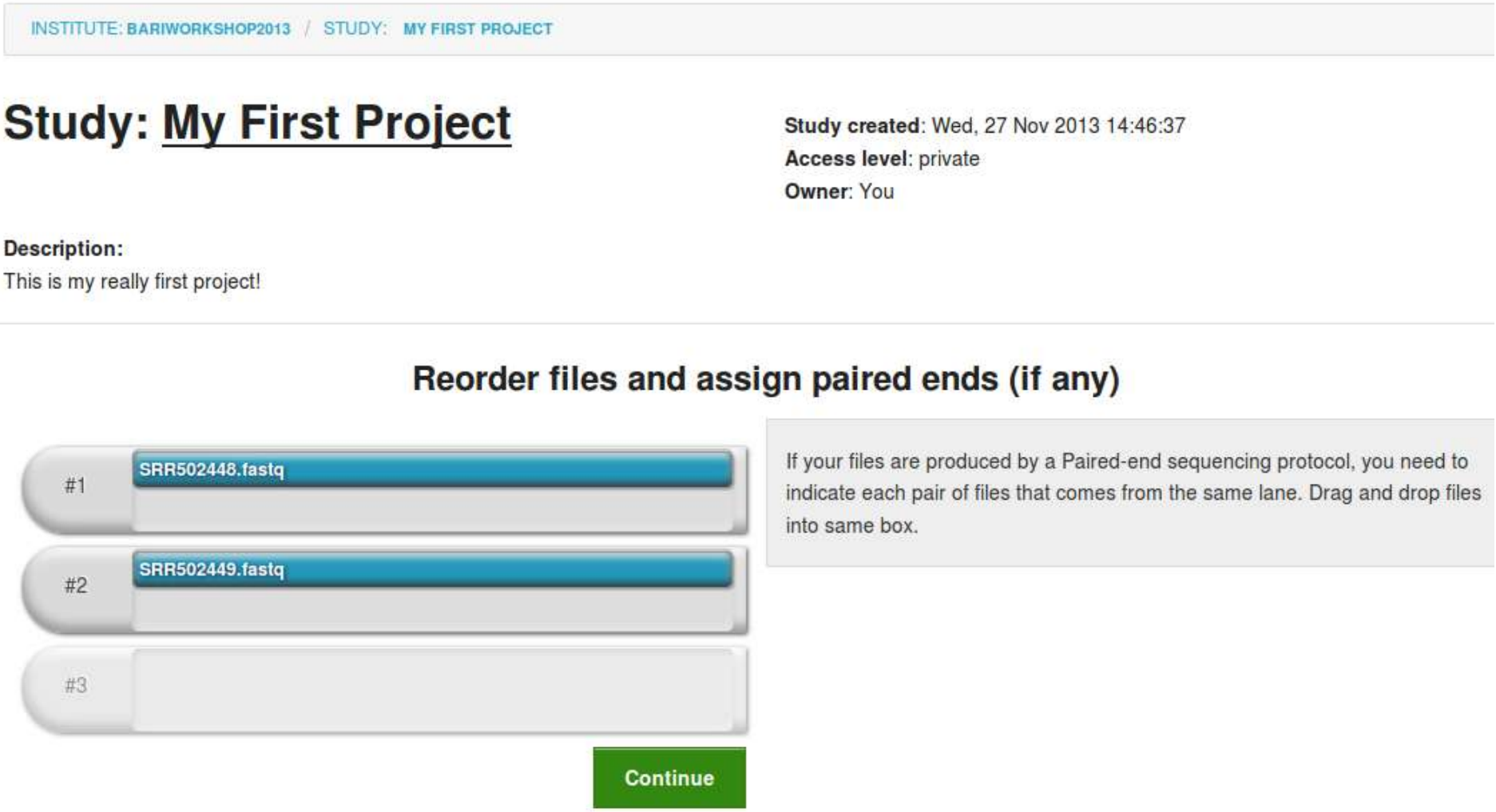


Preprocessing - decompression

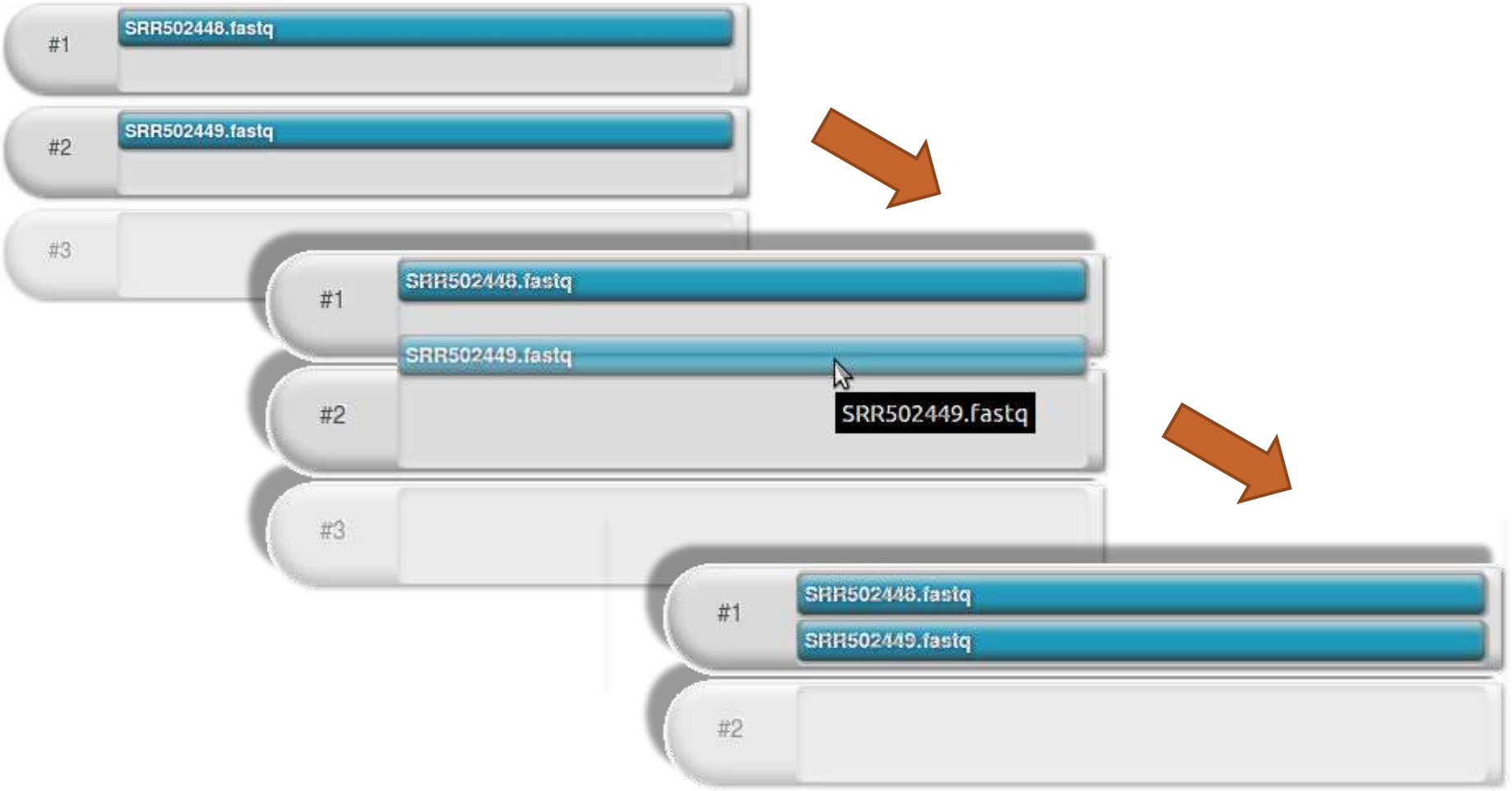
You will be notified by email once completed



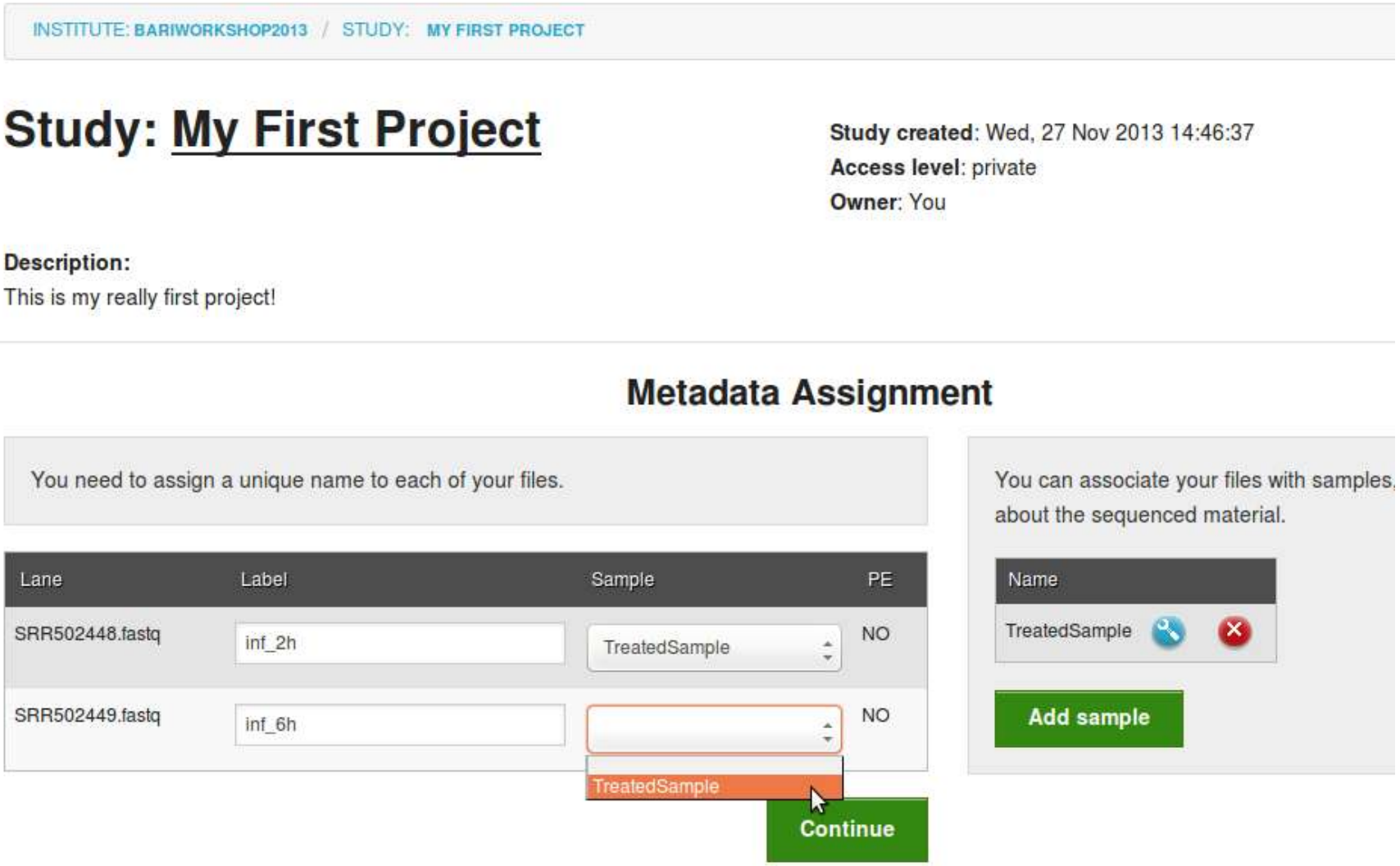
Preprocessing - lane pairing



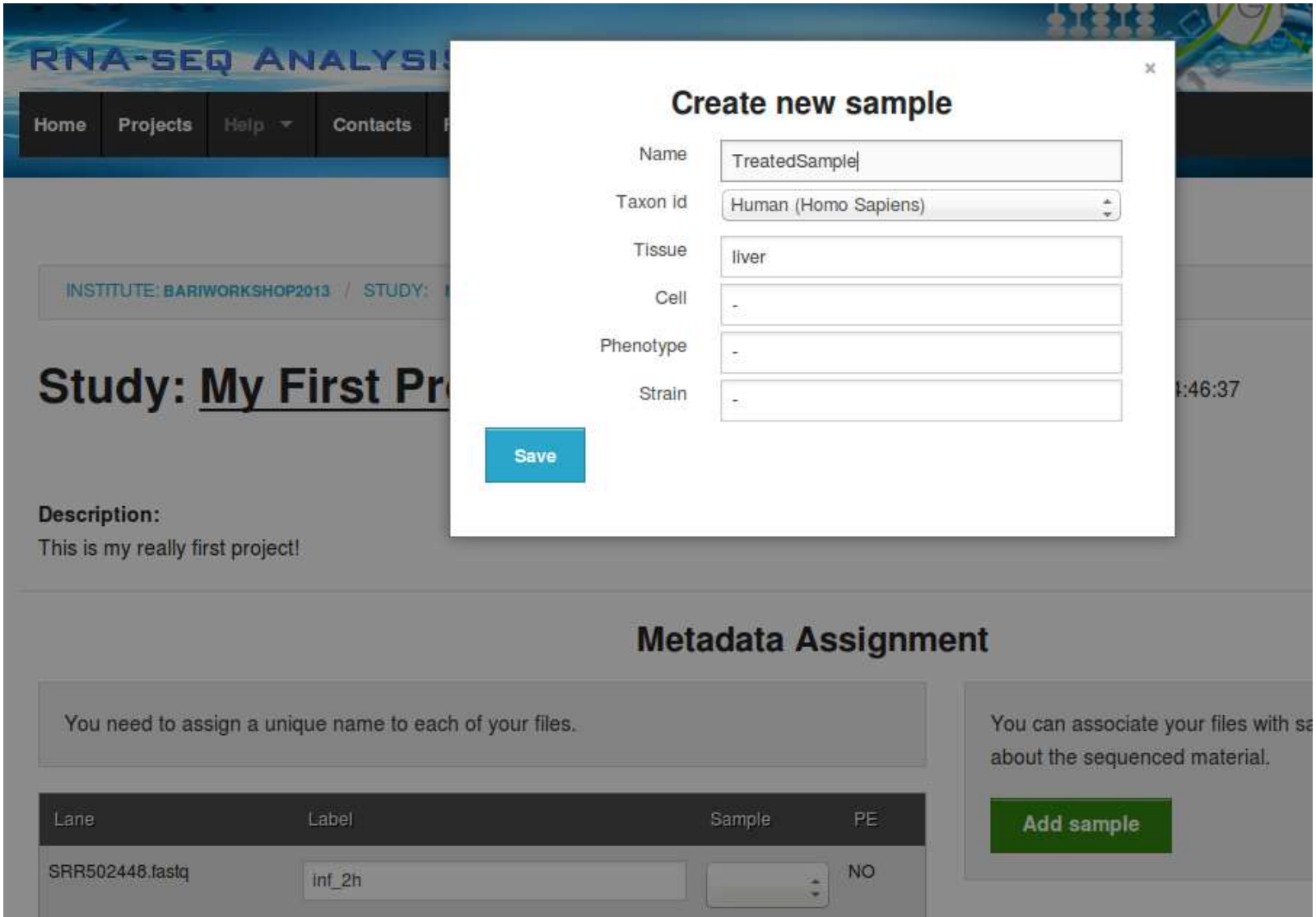
Preprocessing - lane pairing



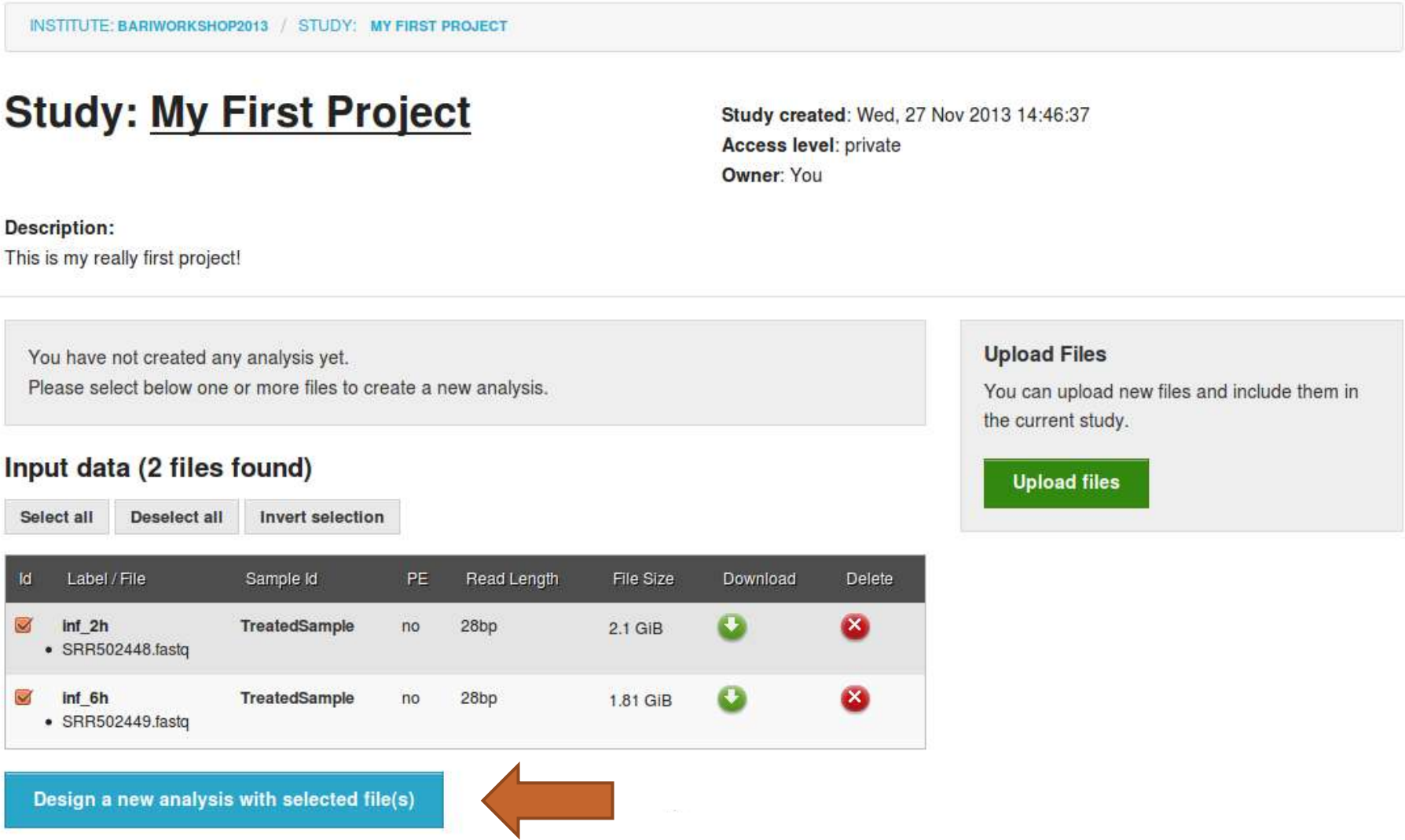
Preprocessing - metadata



Preprocessing - metadata



Preprocessing - files uploaded



Analysis Parameters

Pipeline parameters *(You can safely use defaults)*

Common parameters

--mate-inner-dist	200	*This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200.* (default: 200)
Database	hg18	*Choose the database for the alignment*
Reference-GTF	hg18_gtf	*Reference GTF used to guide assembly*
--library-type	fr-unstranded	***fr-unstranded**: Reads from the left-most end of the fragment map to the transcript strand, and the right-most end maps to the opposite strand.* ***fr-firststrand**: The right-most end of the fragment is the first sequenced.* ***fr-secondstrand**: The left-most end of the fragment is the first sequenced.*
Search-novel-junctions	☑	*Enable search of novel junctions* (default: on)
Search-cassette-exons	☐	*Enable search of cassette exons* (default: 0)
Search-polya	☑	*Enable search of polyA sites (requires junction detection as preliminary step)* (default: on)
Fusion-search	☐	*Enable Fusion Transcripts discovery*

Quality check and filtering

Length	70	*The cut-off value for percentage of read length that should be of given quality* (default: 70)

Analysis Parameters

Quality check and filtering

Length	70	*The cut-off value for percentage of read length that should be of given quality* (default: 70)
Quality	20	*The cut-off value for PHRED quality score for high-quality filtering* (default: 20)
PrimerAdaptorFilter	Browse... No file selected.	*File for user defined primer/adaptor sequences, one per line*
PrimerAdaptorLibrary	-	*Selection of primer/adaptor library (incompatible with PrimerAdaptorFilter)*

Transcript assembly and aboundancy estimation

Novel-Transcripts	☐	*If selected, output will include all reference transcripts as well as any novel genes and isoforms that are assembled.* *If unselected, it will not assemble novel transcripts, and the program will ignore alignments not structurally compatible with any reference transcript.* (default: 0)

Genome alignment

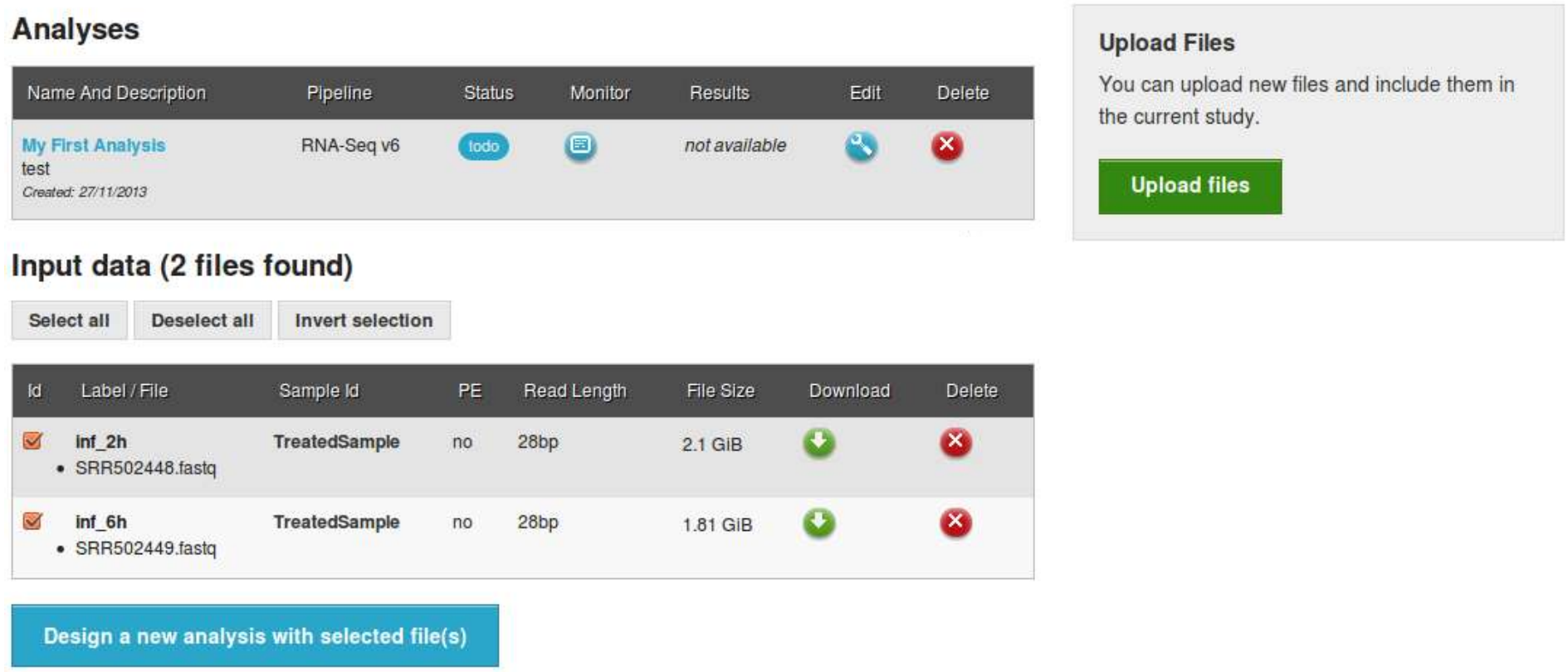
--read-edit-dist	2	*Final read alignments having more than these many edit distance are discarded.* (default: 2)
--read-gap-length	2	*Final read alignments having more than these many total length of gaps are discarded.* (default: 2)
--read-mismatches	2	*Final read alignments having more than these many mismatches are discarded.* (default: 2)

Determination of polyA reads

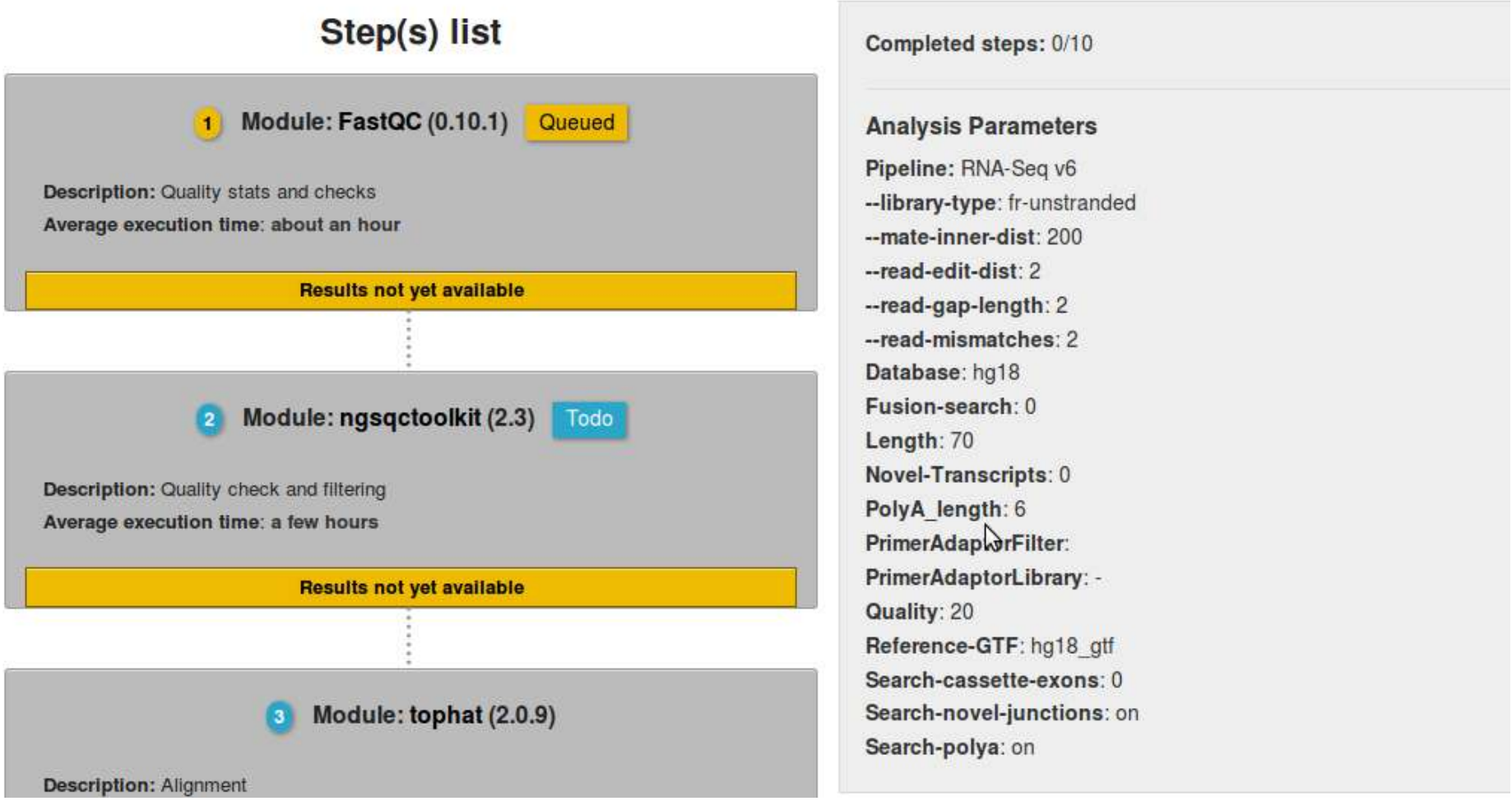
PolyA_length	6	*Length of polyA stretches* (default: 6)

Analysis Started

Your analysis has been queued and each step will be automatically executed



Monitoring an analysis

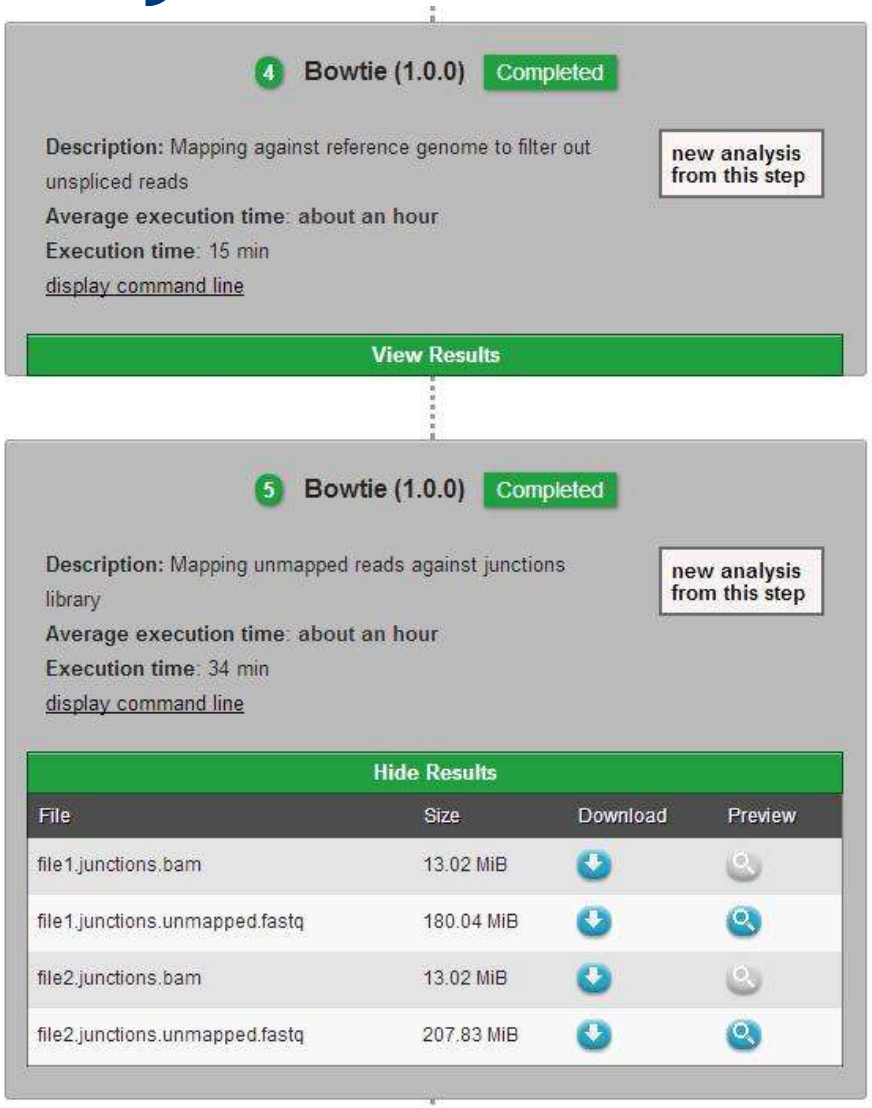


Monitoring an analysis

COMPLETED steps:

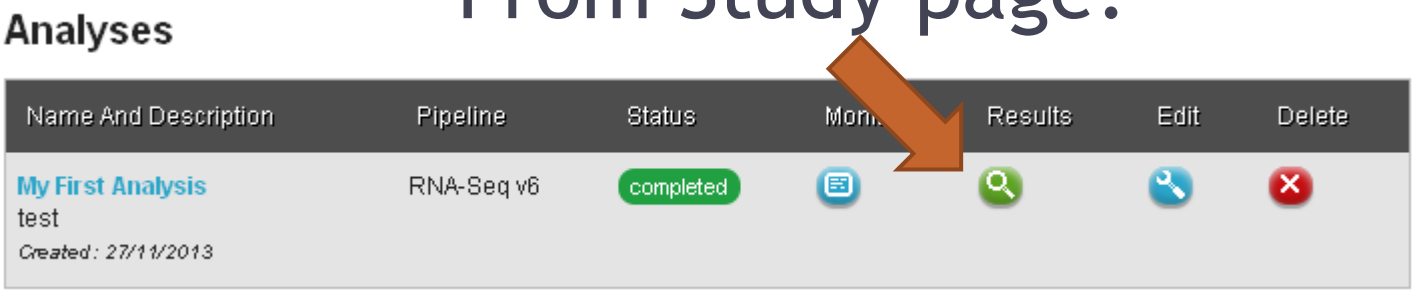
List of files generated

they can be previewed and downloaded

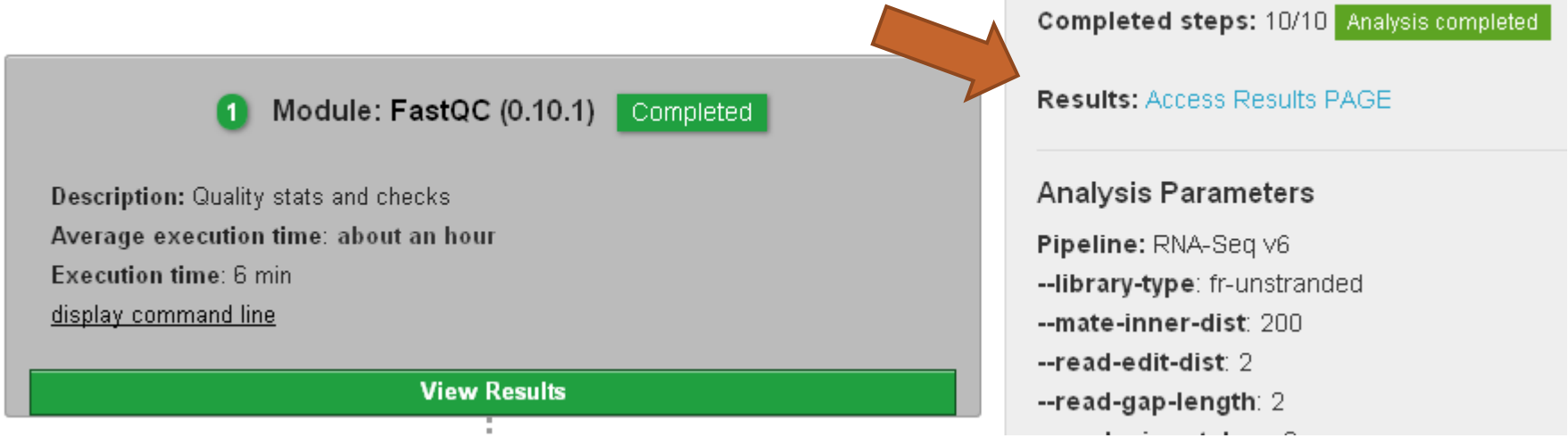


Analysis Completed – Access results

From Study page:



From Monitor Page:

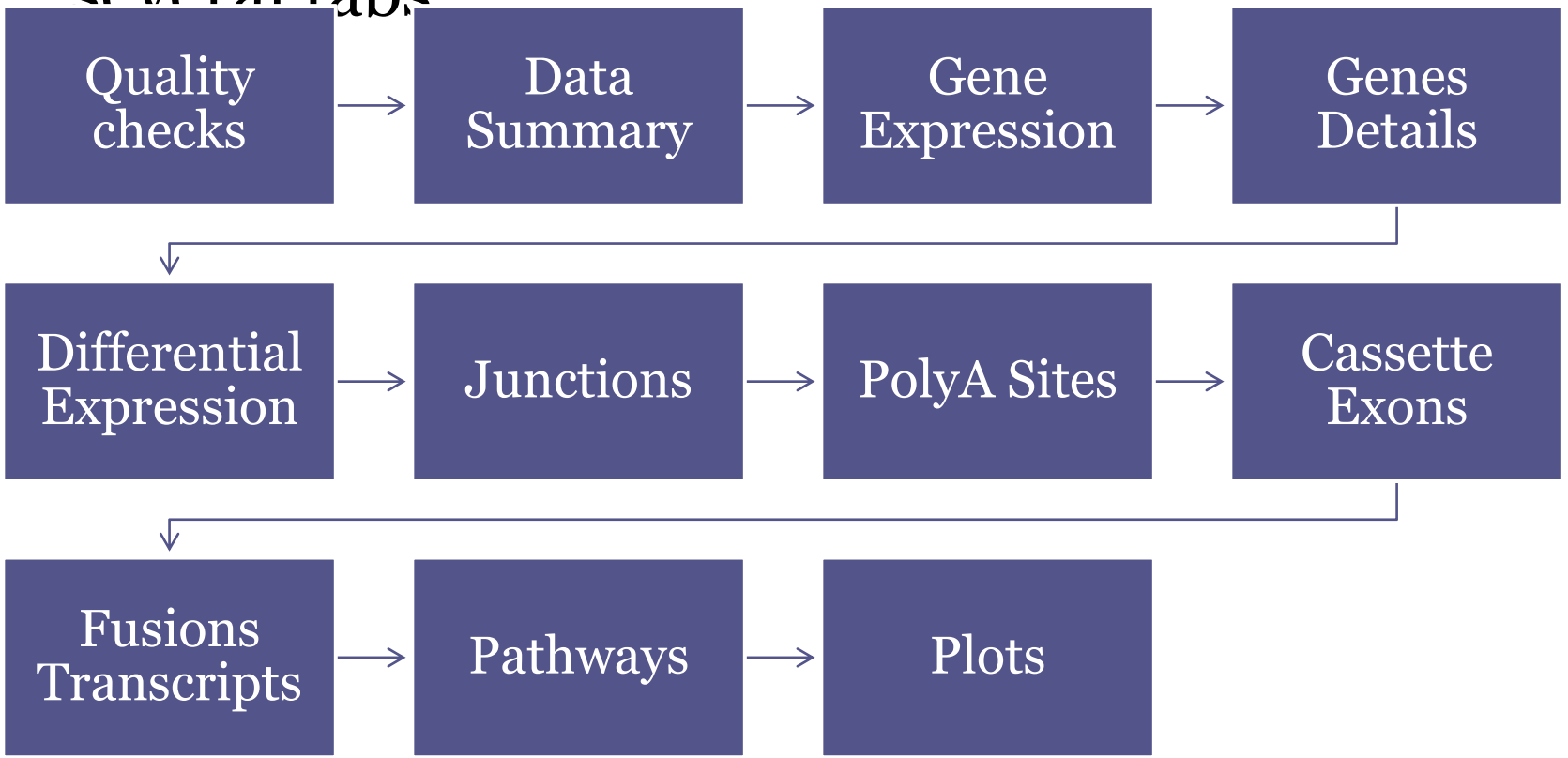


The RAP Analysis Results

Chapter 3

Results visualization

- The user can browse among results through several tabs



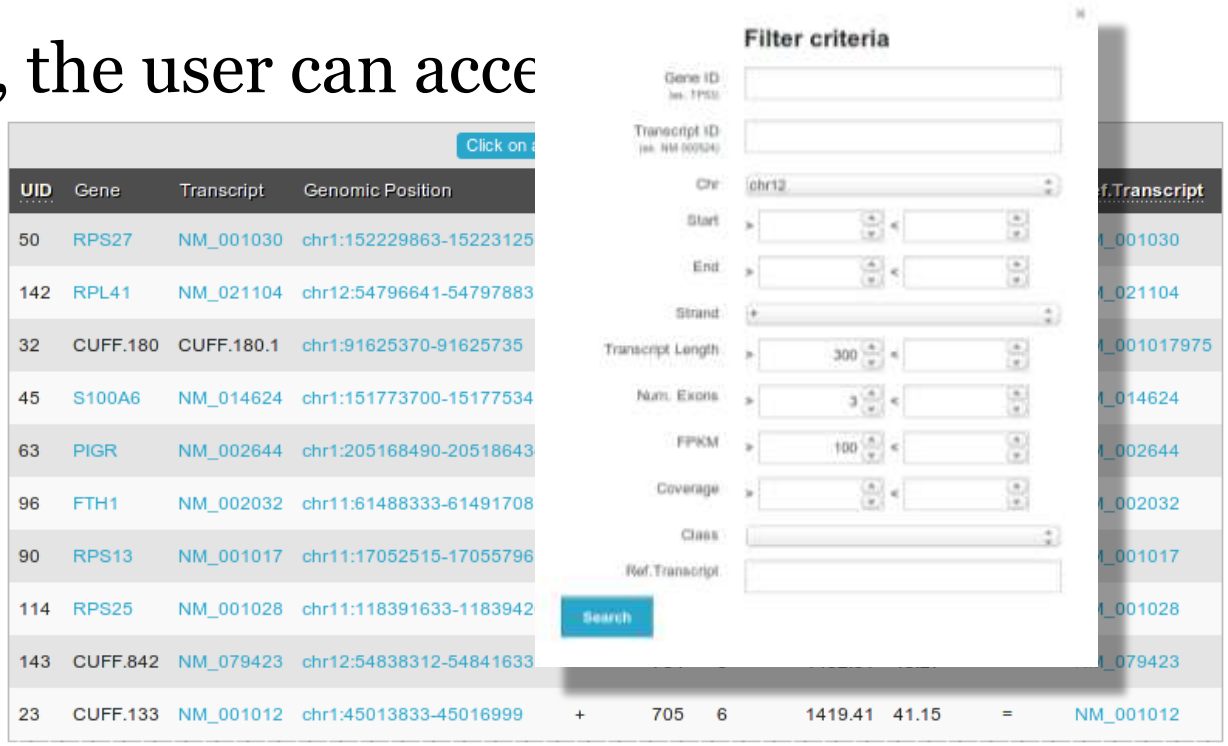
Results visualization

- Embedded display ...mary page

File	Label		Expressed FPKM>0	Expressed FPKM>10	Expressed FPKM>20	Expressed FPKM>100	#HIDATA Loci
1	**Embryonic1**	transcripts	22852	7374	4265	640	
		genes	16963	7180	4355	680	0
2	**Embryonic2**	transcripts	23096	7436	4257	651	
		genes	17160	7196	4363	690	0
3	**Embryonic3**	transcripts	23104	7332	4290	682	
		genes	17160	7126	4364	728	0
4	**Embryonic4**	transcripts	23182	7408	4280	645	
		genes	17223	7203	4376	688	0
5	**Adult1**	transcripts	23989	7198	4148	713	
		genes	17866	6987	4214	754	0
6	**Adult2**	transcripts	23874	7262	4215	725	
		genes	17782	7045	4264	760	0
7	**Adult3**	transcripts	24044	7094	4063	698	
		genes	17891	6884	4128	741	0

Results visualization

- From summary, the user can acce
- to a page of deta

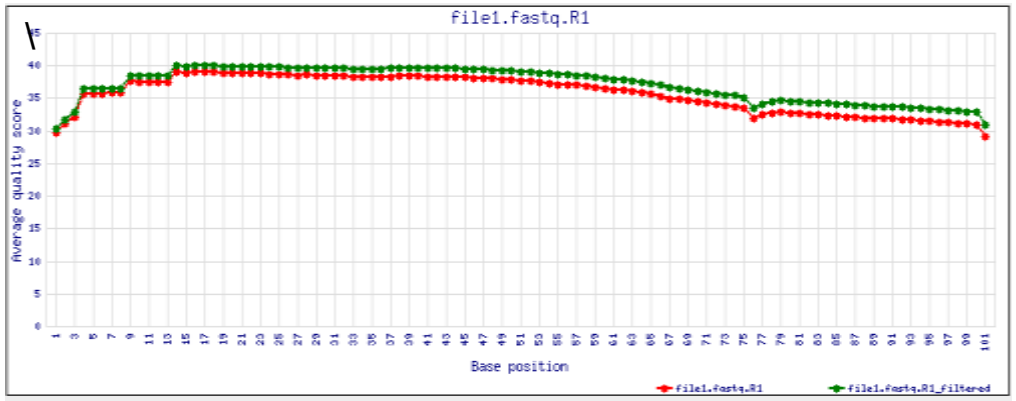


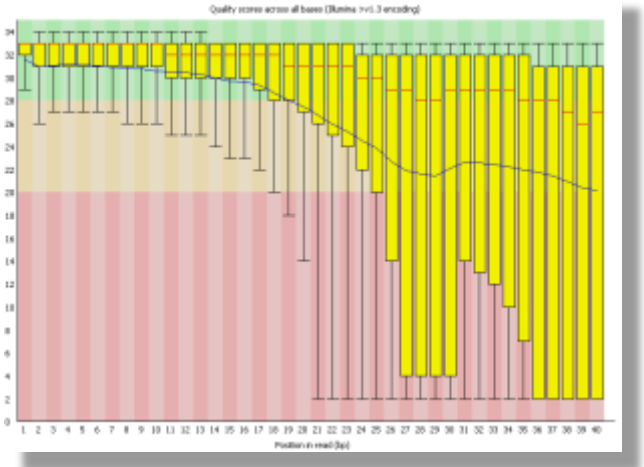
- Each page can b
- filtered by mean[illegible] thresholds
- sorted and
- exported in CSV format via the download engine

Analysis results - Quality checks

- Comprehensive set of quality checks and

File	Label	Checks And Filtering	Basic Statistics	Per Base Sequence Quality	Per Sequence Quality Scores	Per Base Sequence Content	Per Base GC Content	Per Sequence GC Content	Per Base N Content	Sequence Length Distribution	Sequence Duplication Levels	Overrepresented Sequences	Kmer Content	Quality Summary
1	inf_2h		PASS	PASS	PASS	WARN	PASS	WARN	PASS	PASS	PASS	WARN	WARN	QC pair1
2	inf_6h		PASS	PASS	WARN	WARN	PASS	WARN	PASS	PASS	PASS	PASS	WARN	QC pair1

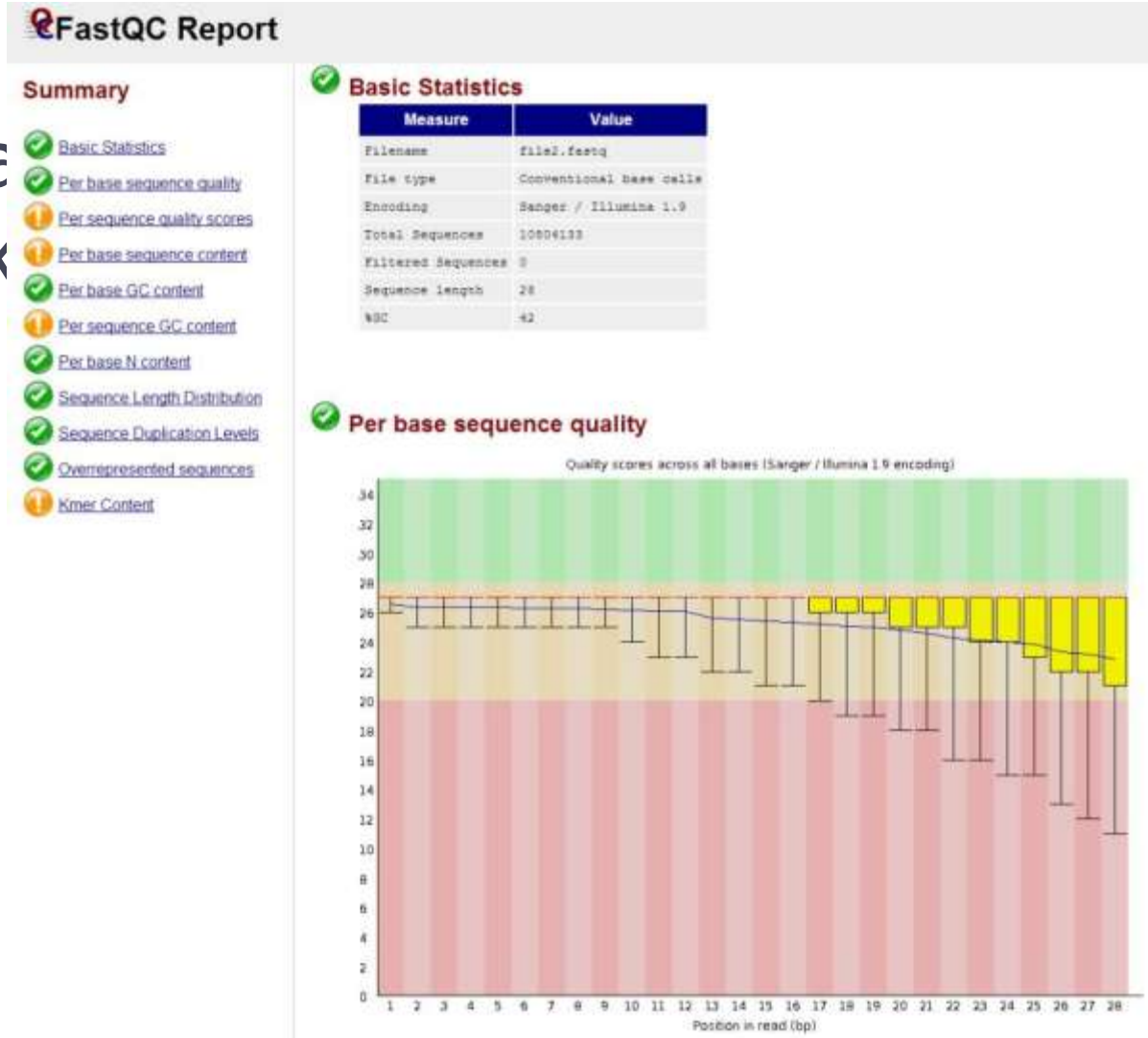




Analysis results - Quality checks

- Details for ea
- quality check

- Graphs
- Stats
- Information



Analysis results – Summary

- Input and high quality reads
- Mapped reads (to the genome)
- Mapped reads (to junctions library)

Summary of RNA-Seq metrics

Organism: Human (Homo Sapiens)
Genome: hg18

File	Label	Input Reads	Tophat	Junctions
1	inf_2h	**Raw**: 12511139 **Filtered**: 12218844	**Overall Mapping**: 11049476 (90.4%)	**Reads**: 48 **Mapped reads**: 48 (100.00%)
2	inf_6h	**Raw**: 10804133 **Filtered**: 10137081	**Overall Mapping**: 8933160 (88.1%)	**Reads**: 62 **Mapped reads**: 62 (100.00%)

Import data into IGV

Analysis results - Gene Expression

Gene and transcript expression summary

Click on the colored-box numbers to open the expression overview

File	Label		Expressed FPKM>0	Expressed FPKM>10	Expressed FPKM>20	Expressed FPKM>100	#HIDATA Loci
1	Embryonic1	transcripts	22852	7374	4265	640	
		genes	16963	7180	4355	680	0
2	Embryonic2	transcripts	23096	7436	4257	651	
		genes	17160	7196	4363	690	0
3	Embryonic3	transcripts	23104	7332	4290	682	
		genes	17160	7126	4364	728	0
4	Embryonic4	transcripts	23182	7408	4280	645	
		genes	17223	7203	4376	688	0
5	Adult1	transcripts	23989	7198	4148	713	
		genes	17866	6987	4214	754	0
6	Adult2	transcripts	23874	7262	4215	725	
		genes	17782	7045	4264	760	0

Analysis results - Gene Expression

Expressed transcripts found for '*inf_2h*'

Filter, browse and download results
Sample: TreatedSample

Page Info

1681	elements
169	pages
10	rows-per-page

Download

List of selected Filters

FPKM	> '20'

Reset Filters | Choose Filters

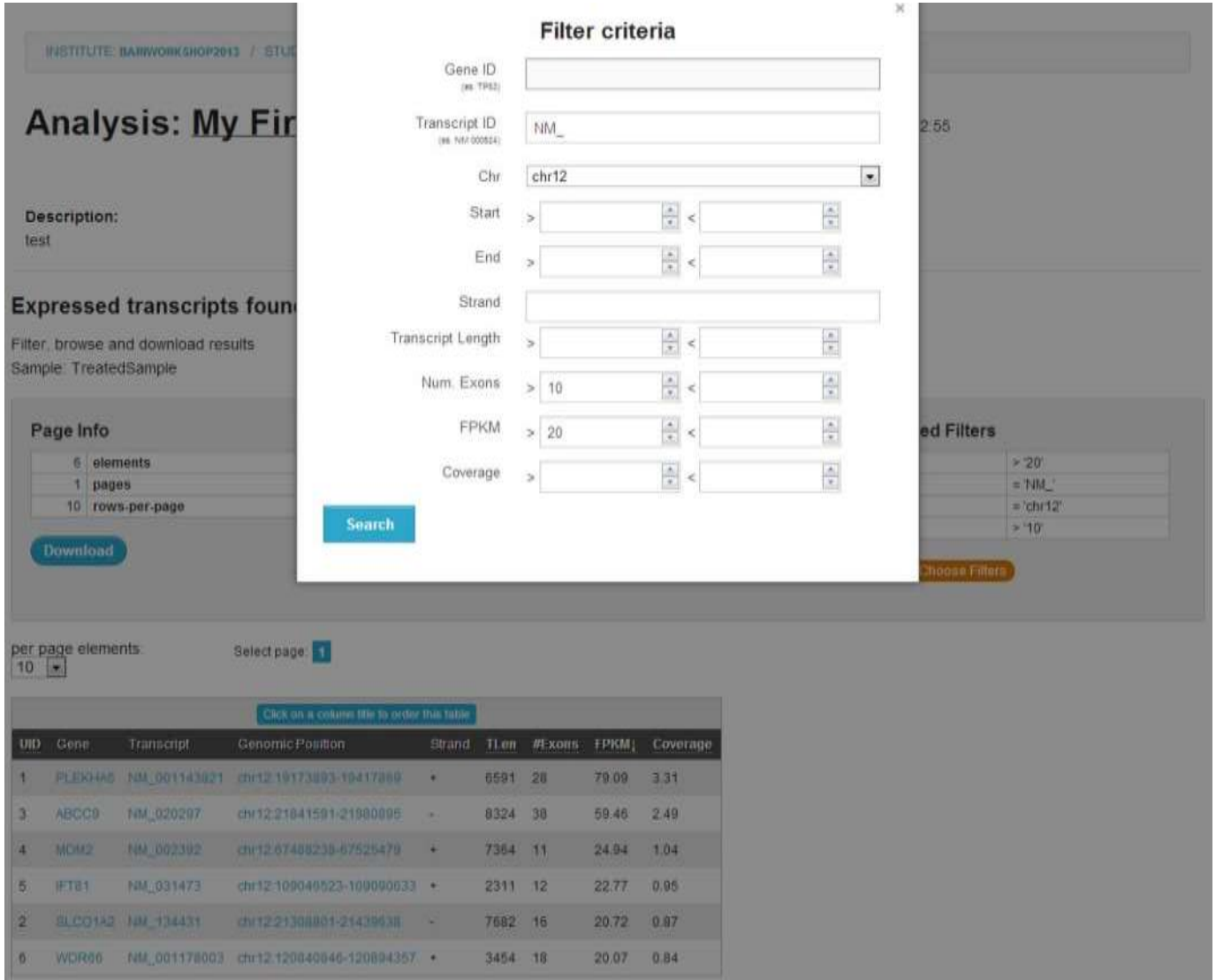
per page elements: 10

Select page: 1 2 3 4 ... »

Click on a column title to order this table

UID	Gene	Transcript	Genomic Position	Strand	TLen	#Exons	FPKM↓	Coverage
1268	MIR4461	NR_039666	chr5:134291628-134291701	+	74	1	237307.93	9918.79
637	MIR548AC	NR_039621	chr17:28547066-28547096	-	31	1	64029.67	2676.26
987	MIR3687	NR_037458	chr21:1678868-1678928	-	61	1	42134.91	1761.12
1206	MIR1267	NR_031671	chr4:177196342-177331125	+	57	3	39547.53	1652.97
672	MIR548O2	NR_039605	chr17:60821546-60847231	-	52	3	34715.01	1450.99
941	MIR663A	NR_030386	chr20:26136822-26136914	-	93	1	16631.98	695.17
1282	MIR548D2	NR_030385	chr5:159002885-159095000	+	81	4	14808.62	618.96
1214	MIR4454	NR_039659	chr5:7322416-7322467	-	52	1	12569.28	525.36
1603	MIR548D1	NR_030382	chr9:123415763-123798763	-	59	4	11998.16	501.49
1207	MIR548AB	NR_039611	chr4:183713766-183720064	-	56	2	11737.12	490.58

Analysis results - Gene Expression



Analysis results - Gene Expression

Page Info

6	elements
1	pages
10	rows-per-page

Download

List of selected Filters

FPKM	> '20'
Transcript	= 'NM_'
Chr	= 'chr12'
#Exons	> '10'

Reset Filters Choose Filters

per page elements: 10

Select page: 1

Click on a column title to order this table

UID	Gene	Transcript	Genomic Position	Strand	TLen	#Exons	FPKM↓	Coverage
1	PLEKHA5	NM_001143821	chr12:19173893-19417869	+	6591	28	79.09	3.31
3	ABCC9	NM_020297	chr12:21841591-21980895	-	8324	38	59.46	2.49
4	MDM2	NM_002392	chr12:67488238-67525479	+	7364	11	24.94	1.04
5	IFT81	NM_031473	chr12:109046523-109090633	+	2311	12	22.77	0.95
2	SLCO1A2	NM_134431	chr12:21308801-21439638	-	7682	16	20.72	0.87
6	WDR66	NM_001178003	chr12:120840846-120894357	+	3454	18	20.07	0.84

Analysis results - Search by Gene

Expression levels for gene: TP53

File	Label	Gene	Trans	Position	Gene FPKM	Transcript FPKM	Coverage
1	inf_2h	TP53	NM_001126115	chr17:7512445-7519536	13	10.21	0.43
1	inf_2h	TP53	NM_001126117	chr17:7512445-7519536	13	0	0
1	inf_2h	TP53	NM_001126116	chr17:7512445-7519536	13	0	0
1	inf_2h	TP53	NM_001126112	chr17:7512445-7531593	13	0.01	0
1	inf_2h	TP53	NM_000546	chr17:7512445-7531593	13	2.77	0.12
1	inf_2h	TP53	NM_001126113	chr17:7512445-7531593	13	0	0
1	inf_2h	TP53	NM_001126114	chr17:7512445-7531593	13	0	0
2	inf_6h	TP53	NM_001126115	chr17:7512445-7519536	14.96	0	0
2	inf_6h	TP53	NM_001126117	chr17:7512445-7519536	14.96	0	0
2	inf_6h	TP53	NM_001126116	chr17:7512445-7519536	14.96	12.76	0.39
2	inf_6h	TP53	NM_001126118	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_001126112	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_000546	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_001126113	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_001126114	chr17:7512445-7531593	14.96	2.19	0.07

Search a gene or transcript in all results

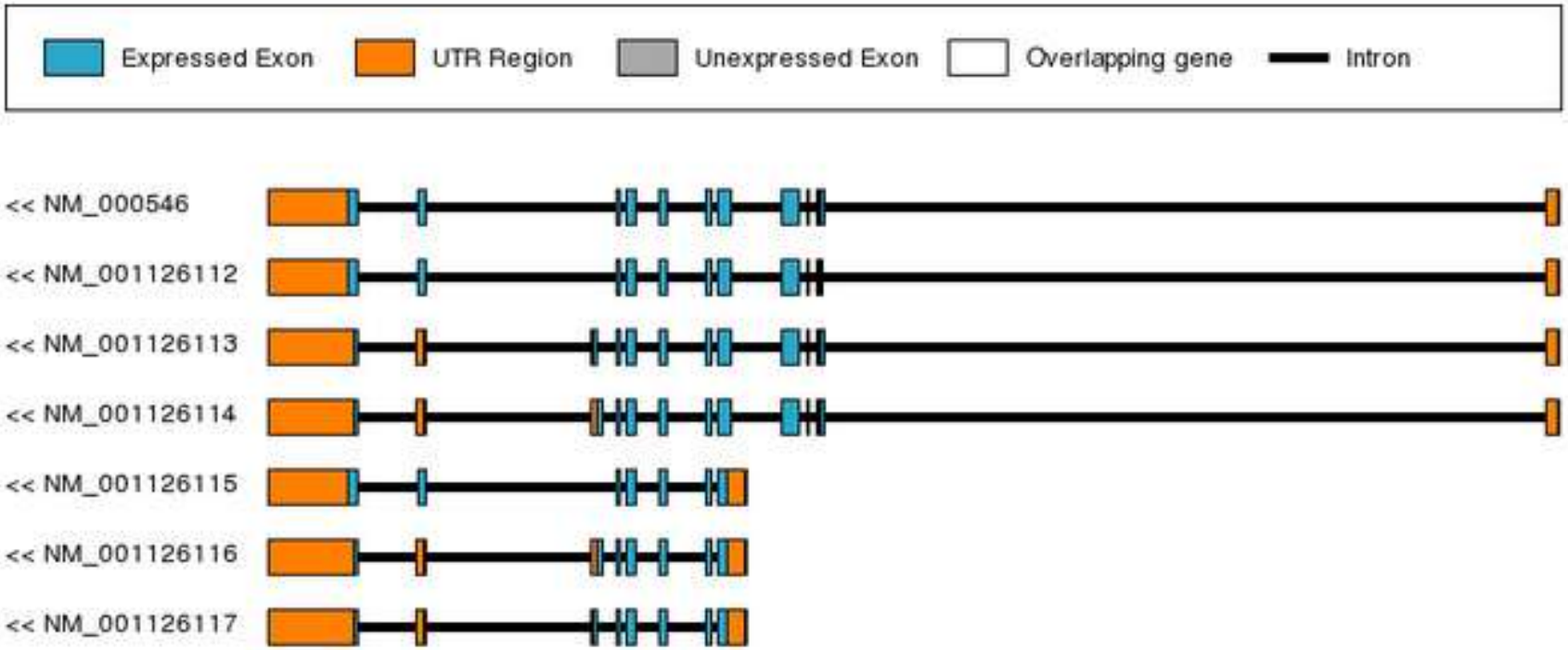
TP53

Check

Analysis results - Gene View

2	inf_6h	TP53	NM_000546	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_001126113	chr17:7512445-7531593	14.96	0	0
2	inf_6h	TP53	NM_001126114	chr17:7512445-7531593	14.96	2.19	0.07

Gene structure view



Analysis results - Junctions

- Novel junctions
- Sele
- skip

File	Label	RefSeq Junctions	Novel Junctions
1	lymph_duod_inf	97095	2588
2	blood_duod_nt	68884	1245
3	lymph_ileum_inf	85910	2106
4	blood_ileum_nt	75358	1564
5	blood_duod_inf	65635	1217
6	lymph_duod_nt	54808	1109
7	blood_ileum_inf	58934	882
8	lymph_ileum_nt	52903	858

Analysis results - Junctions

- Dynamic table
- sort
- filter
- download

Junctions found for dataset '*lymph_duod_inf*'

Filter, browse and download results

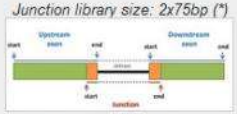
Sample: blood

Page Info

2588	elements
259	pages
10	rows-per-page

Download

Junction library size: 2x75bp (*)



(*) The real size of each junction is the minimum of junction library size and exon length)

List of selected Filters

Junction_Type	= 'novel'

Reset Filters | Choose Filters

per page elements: 10 — Select page: 1 2 3 4 ... »

Click on a column title to order this table

UID	Gene↓	Trans	Chr	Junction	From Exon	To Exon	Junction Type	Num Reads	Gene Coverage
1	0610009O20Rik	NM_026[illegible]	chr11	38257[illegible] 38259971	[illegible] 38257400	[illegible]897 38260056	novel	1	204
2	0610010K14Rik	NM_145758	chr11	70235390 70236098	70235204 70235464	70236024 70236211	novel	3	207
3	0610011F06Rik	NM_026686	chr17	25875652 25876279	25875500 25875726	25876220 25876279	novel	2	369
4	1110001A16Rik	NM_001177402	chr17	78918398 78919833	78918357 78918472	78919759 78919843	novel	3	114
5	1110007C09Rik	NM_026738	chr13	49203718 49205266	49203718 49203789	49205192 49205345	novel	2	131
6	1110008L16Rik	NM_025373	chr12	55304810 55308811	55302637 55304884	55308737 55308869	novel	3	90
7	1110018G07Rik	NM_178065	chr12	84926339 84927943	84926234 84926413	84927869 84927944	novel	1	63
8	1110019D14Rik	NR_045995	chr6	13871569 13895166	13871569 13871627	13895092 13896421	novel	2	12
9	1110032A03Rik	NM_023483	chr9	50764224 50767932	50764224 50764290	50767858 50768152	novel	1	42
11	1110037F02Rik	NM_001081183	chr4	11546226 11549619	11545969 11546300	11549545 11551143	novel	1	76

Analysis results – Cassette Exons

- Cassette exons and other elementary splicing events
- Intron retention
- Alternative 5’ / 3’

Quality checks | Data Summary | Gene expression | Search by Gene | Differential Expression | Isoforms | PolyA Sites | Cassette Exons

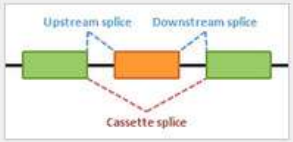
File	Label	Cassette Exons	Constitutive Exons	Intron Retention	Alternative 5'	Alternative 3'	Total
1	Embryonic1	[illegible]	[illegible]	[illegible]	[illegible]	280	83300
2	Embryonic2	2910	85059	142	424	295	88830
3	Embryonic3	2781	[illegible]	132	404	266	83865
4	Embryonic4	2552	84790	144	421	292	88199
5	Adult1	2523	87148	150	442	304	90567
6	Adult2	2867	86680	143	437	310	90437
7	Adult3	2825	84741	133	431	277	88407

Analysis results - Cassette Exons

Page Info

16881	elements
1689	pages
10	rows-per-page

Download



No filter selected

Choose Filters

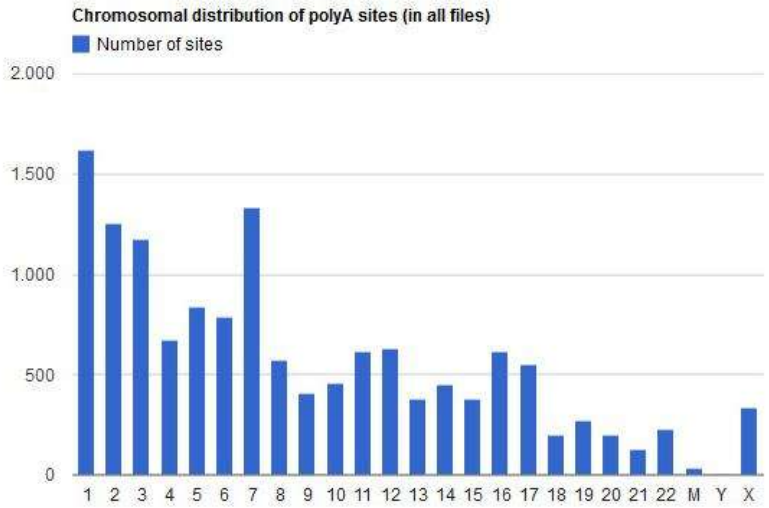
per page elements: 10

Select page: 1 2 3 4 ... »

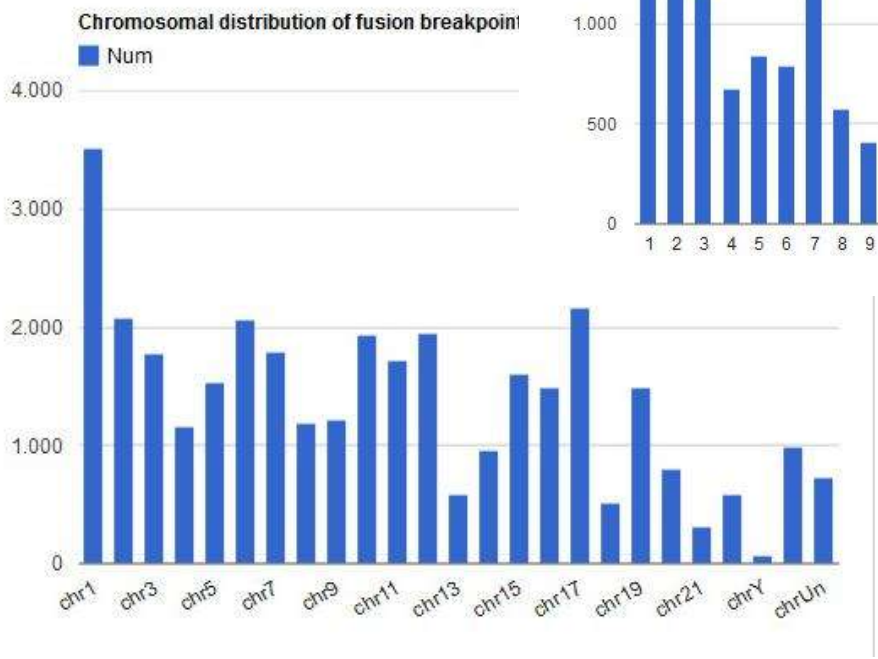
Click on a column title to order this table

UID↓	Gene	Cassette Position	Upstream Exon Position	Downstream Exon Position	Upstream Exon Coverage	Downstream Exon Coverage	Cassette Coverage	AS Type	Inclusion Frequency
16881		chrX_:32429-32584	chrX_:30587-30645	chrX_:39024-39148	168.97	25	101.29	CA	0
16880		chrX_:28975-29113	chrX_:24871-24962	chrX_:30587-30645	167.03	89.66	142.03	CA	0
16879		chrX_:21627-21695	chrX_:14074-14167	chrX_:24871-24962	177.42	27.47	261.76	CA	0
16878	EG621083	chrX:99063364-99063446	chrX:99060727-99060799	chrX:99064393-99064457	52.78	85.94	101.22	CA	0
16877	Cxcr3	chrX:98940786-98940829	chrX:98940121-98940200	chrX:98942777-98942906	155.7	128.68	325.58	CA	0
16876	Cxcr3	chrX:98940121-98940200	chrX:98939772-98939901	chrX:98940786-98940829	105.43	265.12	244.3	CA	0
16875	Taf1	chrX:98763704-98763776	chrX:98732364-98732466	chrX:98780955-98781017	29.41	19.35	50	CA	0
16874	Srpx	chrX:9657040-9657164	chrX:9656379-9656457	chrX:9660979-9661021	326.92	95.24	205.65	CA	0
16873	Yipf6	chrX:96163749-96163910	chrX:96163366-96163497	chrX:96165374-96165467	694.66	1046.24	988.82	CA	0
16872	Yipf6	chrX:96160087-96160251	chrX:96157846-96157943	chrX:96163366-96163497	580.41	901.53	850	CA	0

PolyA sites and fusion transcripts

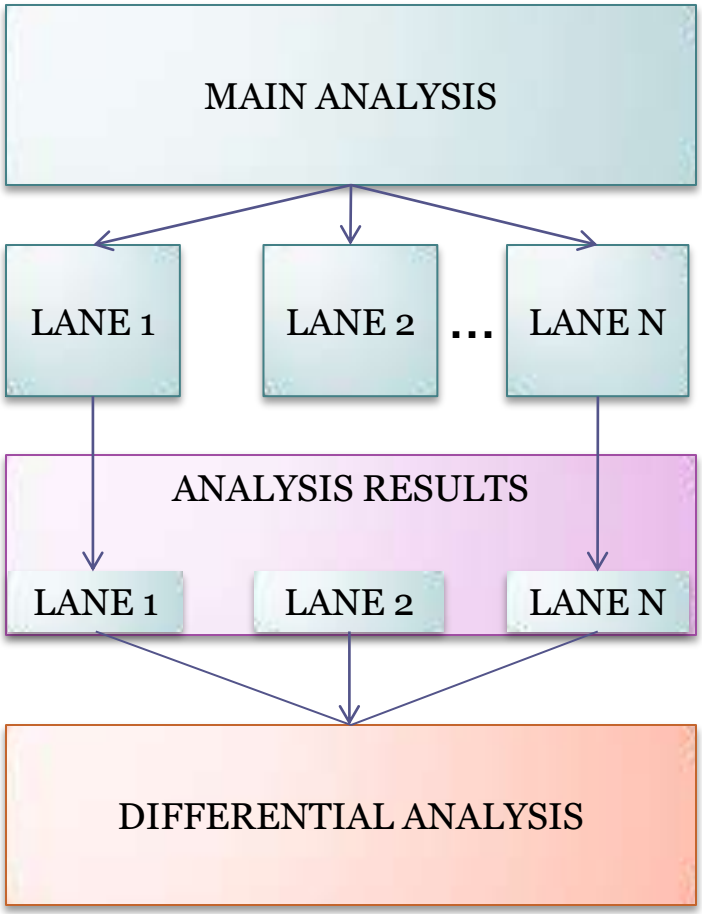


File	Label	Num.sites
1	24h_Met_rep1	2204
2	24h_Met_rep2	2359
3	24h_Met_rep3	1907
4	24h_rep1	2441
5	24h_rep2	2583
6	24h_rep3	2543



File	Label	Fusion Points
1	24h_Met_rep1	2871
2	24h_Met_rep2	2962
3	24h_Met_rep3	2825
4	24h_rep1	3366
5	24h_rep2	3436
6	24h_rep3	3166

Differential Analyses



e.g. expressed genes (per lane)

e.g. differentially expressed genes

Analysis results - Diff. Expression

List of submitted differential expression operations

Date	Type	Group + Analysis + File	Results	Status	Delete
Tue, 22 Oct 2013	cuffdiff	View Lanes	Expand results	completed	
Thu, 31 Oct 2013	cuffdiff	Hide Lanes	Expand results	completed	

group: 2	Validation test	Embryonic1	(file 1)
group: 2	Validation test	Embryonic2	(file 2)
group: 1	Validation test	Adult1	(file 5)
group: 1	Validation test	Adult2	(file 6)

Create a new differential expression operation

If you have replicates, assign them to the same group

File	Filename	Label	Paired-End	Group
1	SRR531311.fastq	Embryonic1	no	

Analysis results – Diff. Expression

Result of a DE operation is a matrix

- pair of samples

Each pair can be expanded for details

Analysis results – Diff. Expression

Differentially expressed genes

Filter, browse and download results

Page Info

206	elements
21	pages
10	rows-per-page

Download

List of selected Filters

Sample1	= 'blood_duod_nt'
Sample2	= 'blood_duod_inf'
Significant?	= 'yes'

Reset Filters Choose Filters

per page elements: 10

Select page: 1 2 3 4 ... »

Click on a column title to order this table

UID	Gene↓	Transcript	Genomic Position	Sample1	Sample2	Status	FPKM Sample1	FPKM Sample2	Log2	Test Stat	Pvalue	Qvalue	Significant?
50	1700011H14Rik	NM_025956	chr14:49226358-49245428	blood_duod_nt	blood_duod_inf	OK	0	12.64	0	0	≤0.01	≤0.01	yes
85	2010002M12Rik	NM_053217	chr19:34617050-34640743	blood_duod_nt	blood_duod_inf	OK	0	11.72	0	0	≤0.01	≤0.01	yes
13	2310007B03Rik	NM_001159940	chr1:93151354-93160948	blood_duod_nt	blood_duod_inf	OK	8.11	0	0	0	≤0.01	≤0.01	yes
34	2310007L24Rik	NM_029345	chr11:106374825-106377114	blood_duod_nt	blood_duod_inf	OK	85.25	0	0	0	≤0.01	≤0.01	yes
64	2310042E22Rik	NM_025634	chr16:21152658-21153944	blood_duod_nt	blood_duod_inf	OK	0	8.71	0	0	≤0.01	≤0.01	yes
20	2310057J18Rik	NM_026336	chr10:28972287-28986306	blood_duod_nt	blood_duod_inf	OK	0	7.54	0	0	≤0.01	≤0.01	yes
181	4930432K21Rik	NM_029045	chr8:84148037-84172597	blood_duod_nt	blood_duod_inf	OK	3.87	0	0	0	≤0.01	0.05	yes
30	Abhd15	NM_026185	chr11:77515116-77520628	blood_duod_nt	blood_duod_inf	OK	4.31	0	0	0	≤0.01	0.02	yes
119	Acap3	NM_207223	chr4:155891874-155907251	blood_duod_nt	blood_duod_inf	OK	2.88	0	0	0	≤0.01	≤0.01	yes
89	Adamtsl2	NM_029981	chr2:27079380-27108613	blood_duod_nt	blood_duod_inf	OK	0	3.66	0	0	≤0.01	≤0.01	yes

Analysis results - Diff. Expression

List of submitted differential expression operations

Date	Type	Group + Analysis + File	Results	Status	Delete
Tue, 22 Oct 2013	cuffdiff	View Lanes	Expand results	completed	
Thu, 31 Oct 2013	cuffdiff	Hide Lanes	Expand results	completed	

group: 2	Validation test	Embryonic1	(file 1)
group: 2	Validation test	Embryonic2	(file 2)
group: 1	Validation test	Adult1	(file 5)
group: 1	Validation test	Adult2	(file 6)

Create a new differential expression operation

If you have replicates, assign them to the same group

File	Filename	Label	Paired-End	Group
1	SRR502448.fastq	inf_2h	no	group 1
2	SRR502449.fastq	inf_6h	no	

group 1
group 2

Select 2 or more groups to activate this ope

Calculate

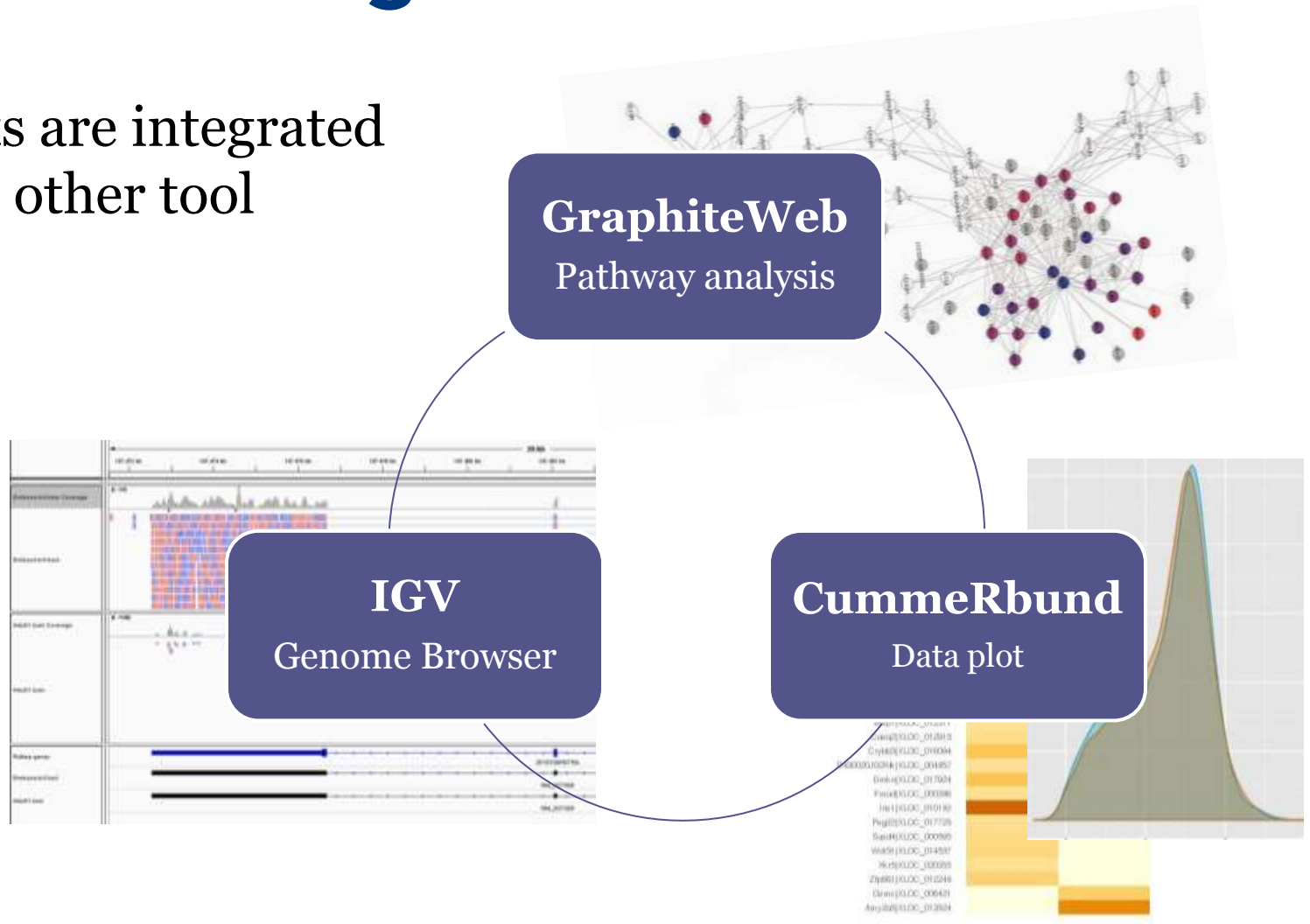
DE operations can be requested as a further analysis

Available differential analyses

- Differential transcript expression
  - Cuffdiff
- Differential gene expression
  - DESeq
- Differential exon inclusion ratio
  - Custom, chi-square
- Differential splicing junction usage
  - DESeq
- Differential Polyadenilation sites usage
  - DESeq

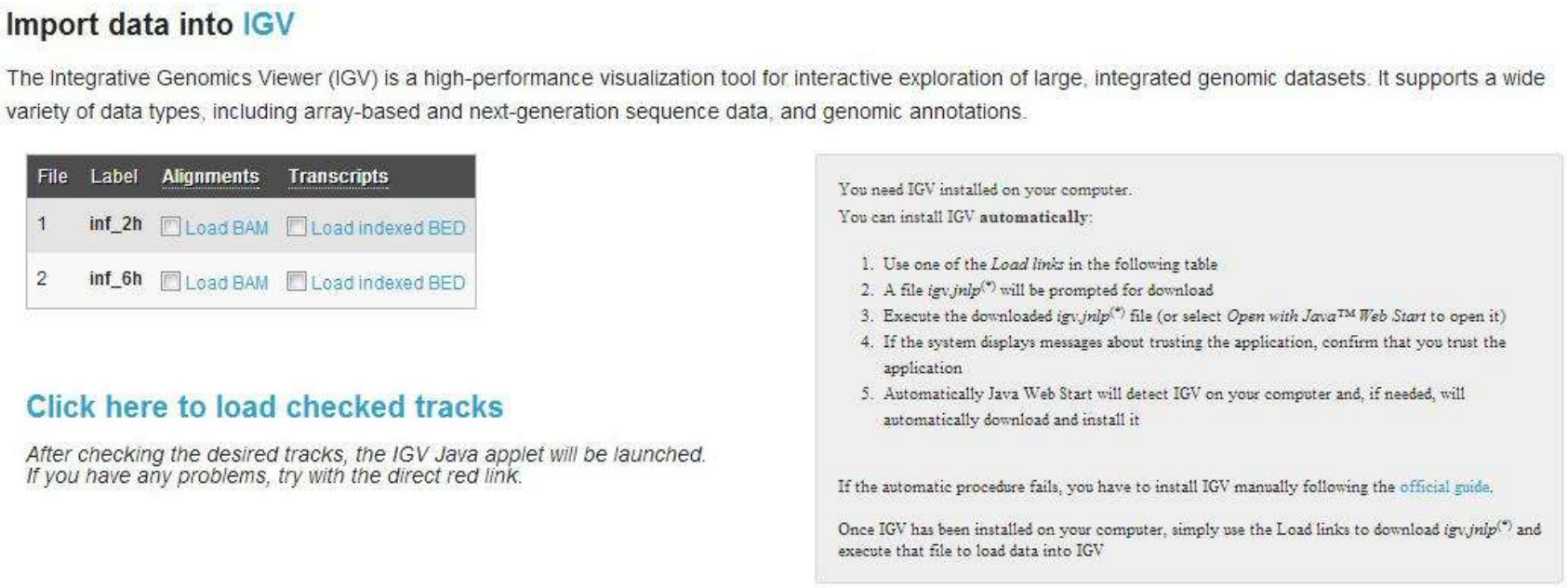
Results integration

Results are integrated with other tool



Integration with IGV

Alignment results and assembled transcripts can be imported into the Integrative Genome Viewer (IGV)



Import data into IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

File	Label	Alignments	Transcripts
1	inf_2h	Load BAM	Load indexed BED
2	inf_6h	Load BAM	Load indexed BED

Click here to load checked tracks

After checking the desired tracks, the IGV Java applet will be launched. If you have any problems, try with the direct red link.

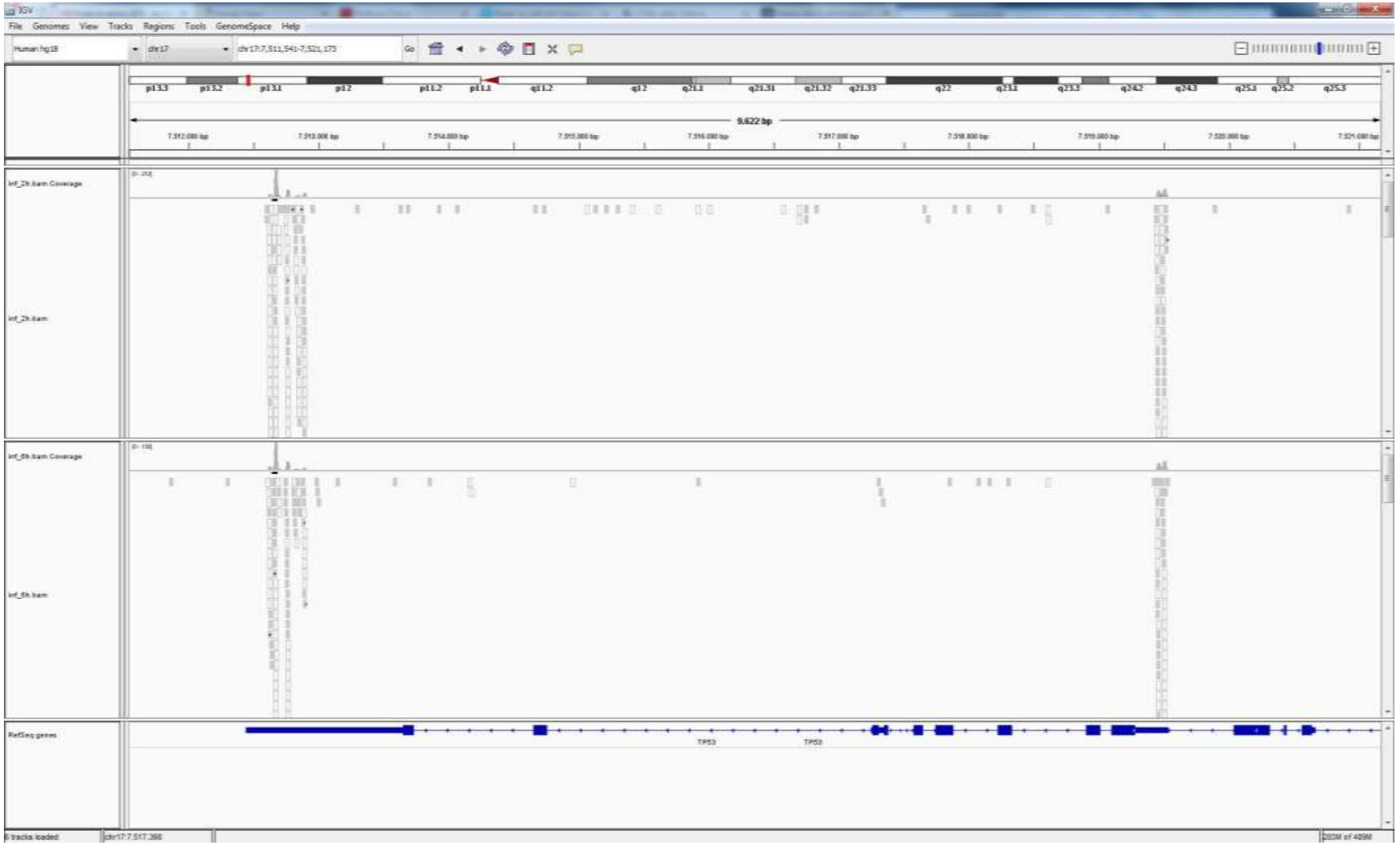
You need IGV installed on your computer.
You can install IGV **automatically**:

1. Use one of the *Load links* in the following table
2. A file *igv.jnlp*(*) will be prompted for download
3. Execute the downloaded *igv.jnlp*(*) file (or select *Open with Java™ Web Start* to open it)
4. If the system displays messages about trusting the application, confirm that you trust the application
5. Automatically Java Web Start will detect IGV on your computer and, if needed, will automatically download and install it

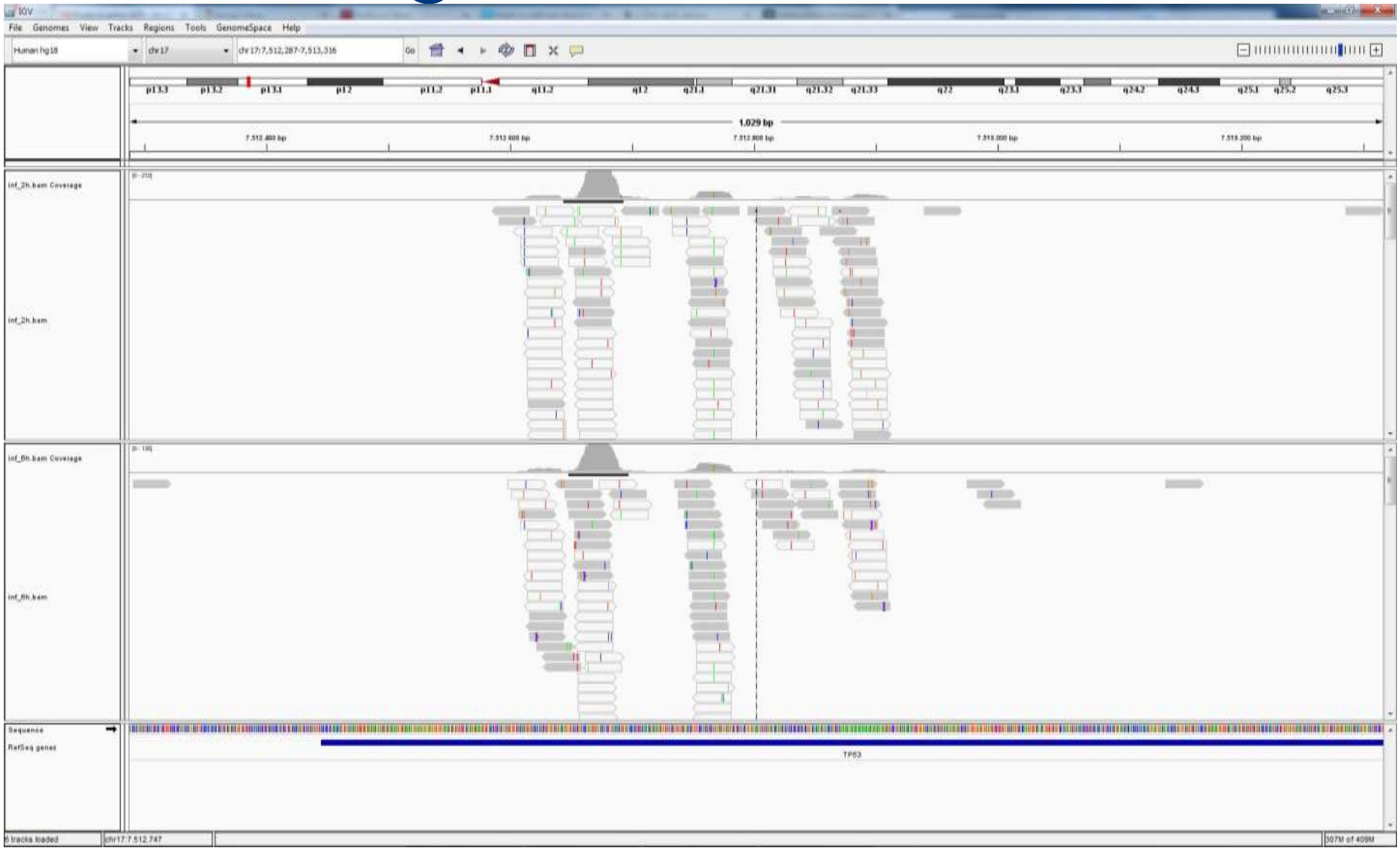
If the automatic procedure fails, you have to install IGV manually following the official guide.

Once IGV has been installed on your computer, simply use the Load links to download *igv.jnlp*(*) and execute that file to load data into IGV

IGV - Integrative Genome Viewer



IGV - Integrative Genome Viewer



Integration with GraphiteWEB

Pathway analyses of gene expression data

Pathway analyses are performed by Graphite Web, a public web server for the analysis and visualization of biological pathways. It supports five different gene set analyses, three species and two pathway databases. Graphite Web has a powerful visualization that makes results interpretation easily accessible to the user.
A tutorial for Graphite Web and its analyses is available here.

Nucleic Acids Res. 2013 Jul;41(Web Server issue):W89-97. doi: 10.1093/nar/gkt386. Epub 2013 May 10.

Graphite Web: Web tool for gene set analysis exploiting pathway topology.

Sales G, Calura E, Martini P, Romualdi C.

PMID 23666626

Your dataset has been succefully imported. Now customize your GraphiteWeb analysis

GraphiteWEB configuration

Please choose one of the following analysis methods:
Hypergeometric test

Choose a pathway database:
KEGG

Hypergeometric test estimates the probability of observing a given number of genes from a pathway among the selected differentially expressed genes.

GSEA compares the distribution of pathway genes adjusted for their correlation structure to that of all the genes provided by the user. This method only works from Expression Analysis results.

SPIA captures several aspects of the data combining the fold change of

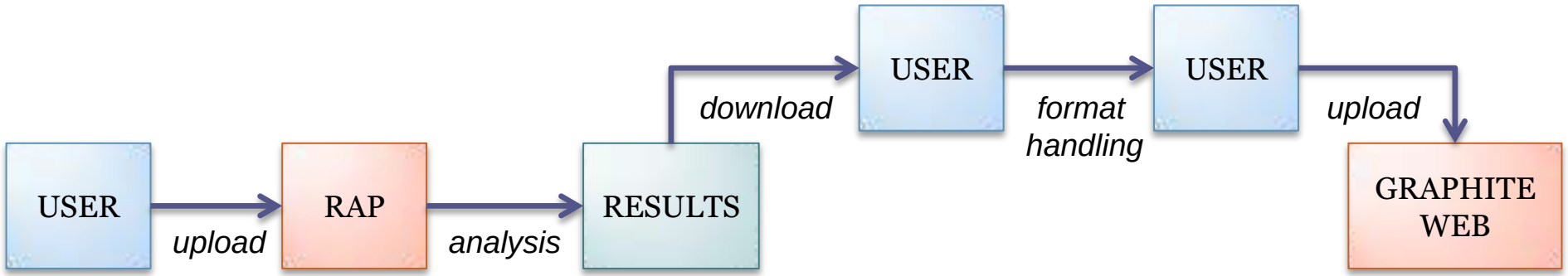
Pathway Analysis with GraphiteWEB

GraphiteWeb

- http://graphiteweb.bio.unipd.it

A fruitful collaboration with Chiara Romualdi

- University of Padova



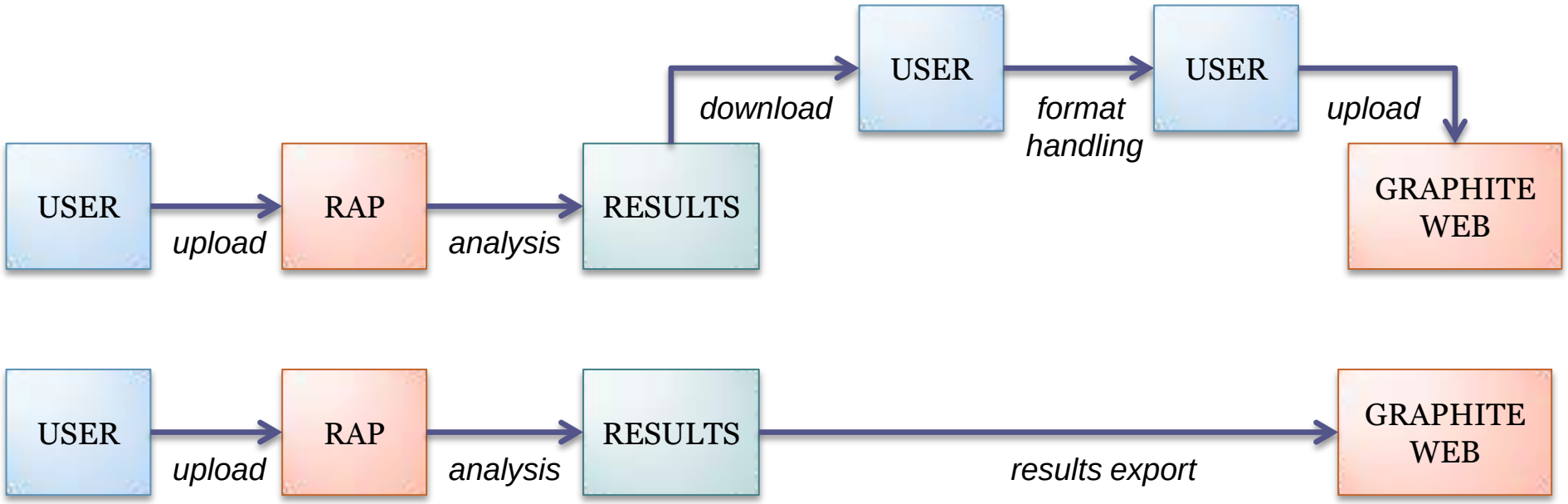
Pathway Analysis with GraphiteWEB

GraphiteWeb

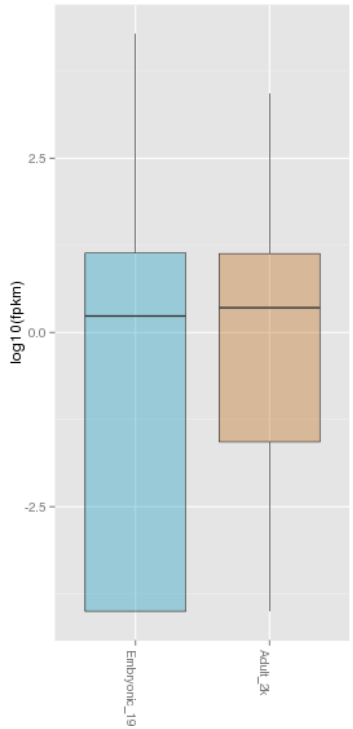
- http://graphiteweb.bio.unipd.it

A fruitful collaboration with Chiara Romualdi

- University of Padova



Integration with CummbeRbund



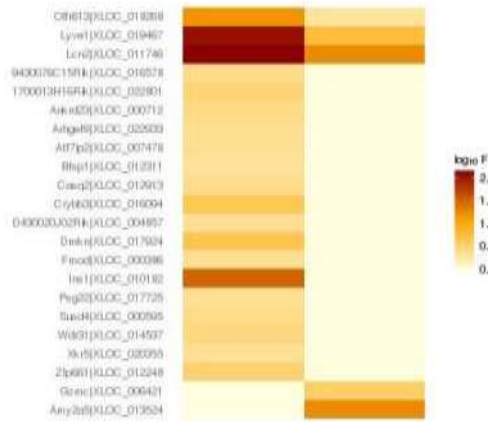
Choose the desired plot:

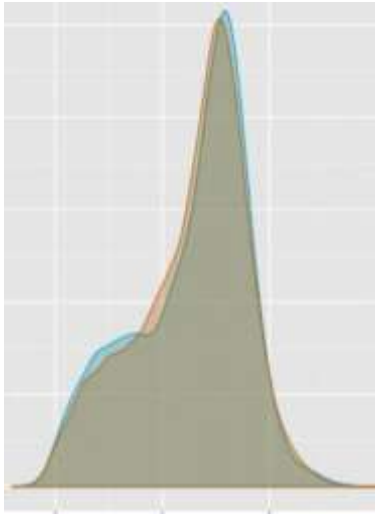
Scatter Plot

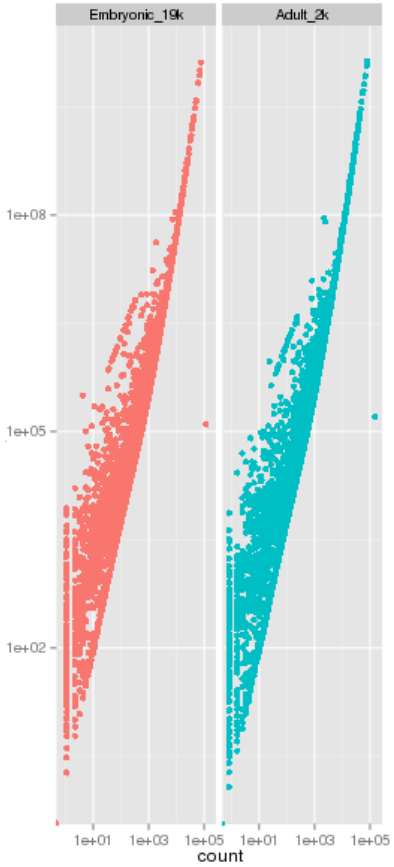
Draw plots per conditons or replicates

Draw per condition

Send request







Thanks for your attention

RAP	http://bioinformatics.cineca.it/rap
Mattia D'Antonio	m.dantonio@cineca.it
Technical support	hpc-service-bio@cineca.it