NGS 雲端分析簡介

李彥樑 (Jack)
威健生技
Welgene Biotech.

Welgene Biotech. Co. Ltd.
http://www.welgene.com.tw

Huge Amount of Sequencing Data

>100G Bases per run per machine

illumina

APPLICATIONS SYSTEMS SERVICES SCIENCE

Systems / HiSeq 2000

HiSeq 2000

The HiSeq 2000 seque
and a breakthrough us
proven and widely-adc
sequencing by synthes
engineering, HiSeq 20
sequencing output and
interaction design feat
set a new standard for

Workflow & Specs

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gb

Library Type	Read Length	Days/Run	Total Tags/Run	Mappable Data
Mate-Paired	2 x 35 bp	8–9	>1.4 B	50–70 GB
	2 x 50 bp	12–16	>1.4 B	80–100 GB
Paired-End	50 bp x 25 bp	11–13	>1.4 B	55–70 GB
Fragment	1 x 35 bp	3.5–4.5	>700 M	25–35 GB
	1 x 50 bp	6–8	>700 M	40–50 GB

From www.Illumina.com

From www. http://www3.appliedbiosystems.com.com

單一RNA樣品數據運算資源

Example: human RNA-seq data

No. of Reads	×	Read Length	=	No. of Base
75.7 M reads	×	50 base	=	3.79G base

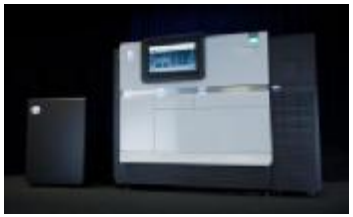
6 CPUs, 12G RAM工作站
1.5天運算時間

~3 G Bytes mapped BAM file = ~9 G Bytes 50SE csfasta file

Sequencing is only the beginning

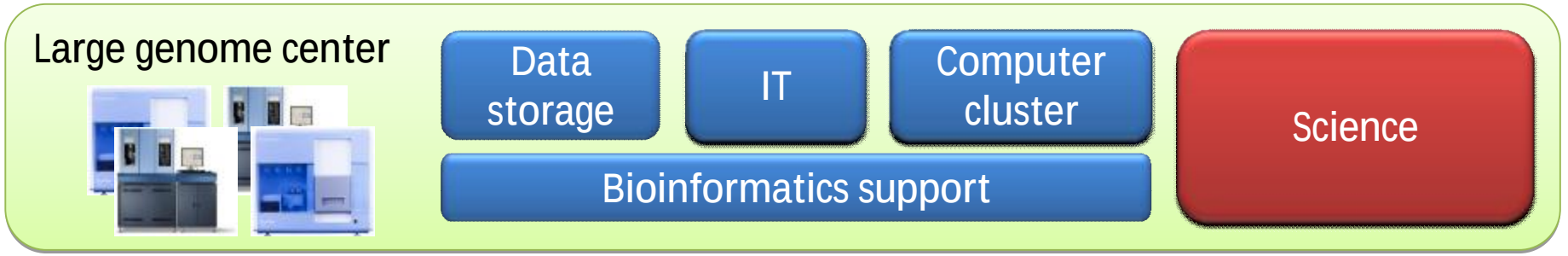






Compute resources?

Sequence Analysis “Gap”



DNAnexus

Accessible to everyone: easy, cost-effective

Address NGS data challenges

- Millions of sequence need to be processed

Accelerate data analysis for customers

- Web-based (no software)
- Cloud infrastructure (no hardware)

A Friendly, Easy, Economic NGS Analysis Interface

Research lab

GATCATGTCACTATACG

Sequencing Service

DNAnexus

Read mapping
Visualization
RNA-seq
ChIP-seq
Variant discovery
Methyl-seq

Future Applications:

- Metagenomics
- De novo assembly, annotation
- Data access APIs
- Custom cloud analyses

Re-sequencing Based Applications

- Genome/transcriptome mapping
- RNA-seq: expression quantification, 3'-end quantification and discovery
- ChIP-seq: Identification of TF binding sites and broad regional interactions (e.g. histone modifications)
- Tag-based enrichment: general discovery and quantification of enrichment
- HpaII/MspI Methyl-seq: enzyme digest site quantification
- Nucleotide-Level variation analysis: mutation analysis
- Cancer variation analysis : tumor/normal sample comparisons to subtract out germline variants
- Small insertion and deletion detection

使用網頁介面分析NGS數據

Cloud Computing Service

維基百科 自由的百科全書 中描述的雲端雲端運算服務特徵：

- 基於虛擬化技術快速部署資源或獲得服務
- 實作動態的、可伸縮的擴充功能
- 按需求提供資源、按使用量付費
- 透過互聯網提供、面向海量資訊處理
- 使用者可以方便地參與
- 形態靈活，聚散自如
- 減少使用者終端的處理負擔
- 降低了使用者對於IT專業知識的依賴

取自http://ithelp.ithome.com.tw/question/10009336

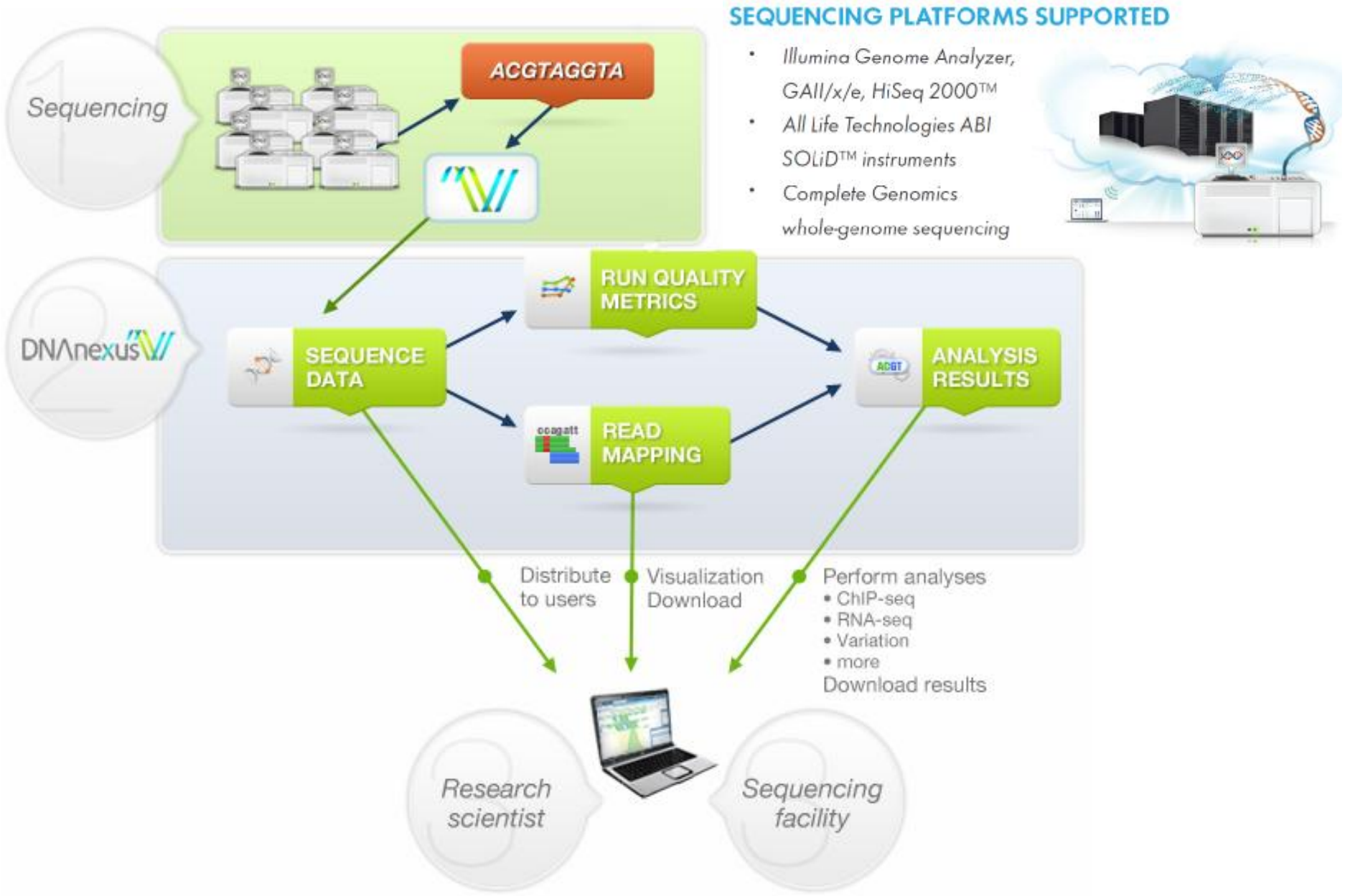
取自http://www.carlosblanco.com/2009/05/14/cloud-computing/

淺談雲端運算 (Cloud Computing)

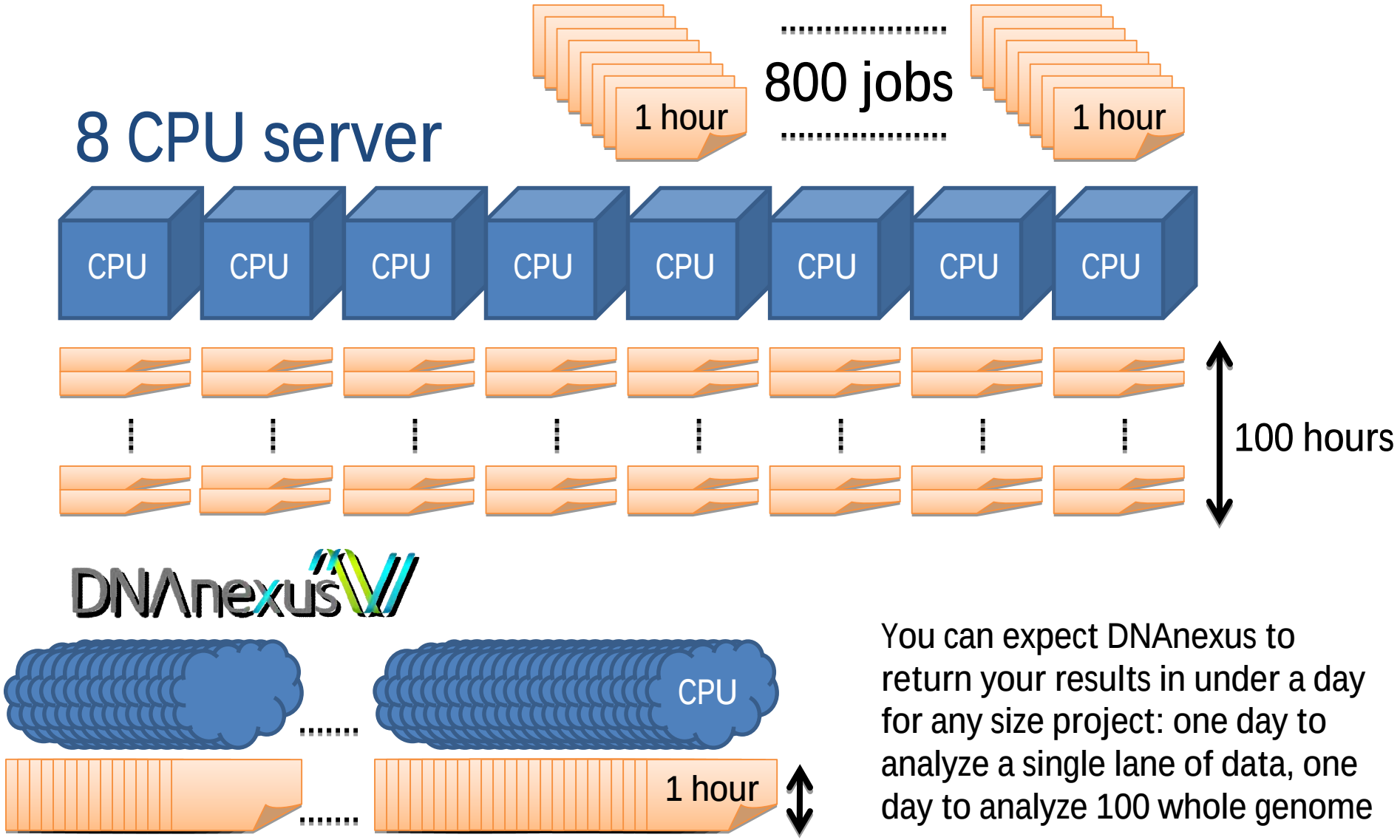
作者：黃重憲 / 臺灣大學電機資訊學院資訊工程系

「雲端運算」=「網路」=「網路運算」。「雲端運算」不是「新技術」或「技術」。「雲端運算」是一種概念，代表的是利用網路使電腦能夠彼此合作或使服務更無遠弗屆。在實現「概念」的過程中，產生出相應的「技術」。

Integrated Solution on Cloud



Compute Resources



You can expect DNAnexus to return your results in under a day for any size project: one day to analyze a single lane of data, one day to analyze 100 whole genome sequences.

Parallel Computing Power Beyond Comparison

With DNAnexus, you can analyze 100 whole human genome sequences in one day.

By building our infrastructure on Amazon EC2 Services, the world's leading cloud computing provider, 100,000s of CPUs and 100s of petabytes of storage are available to you through DNAnexus.

amazon web services Amazon Elastic Compute Cloud (Amazon EC2)

No More Gap

Large genome center

Data storage

IT

Computer cluster

Bioinformatics support

Science

Core sequencing facility

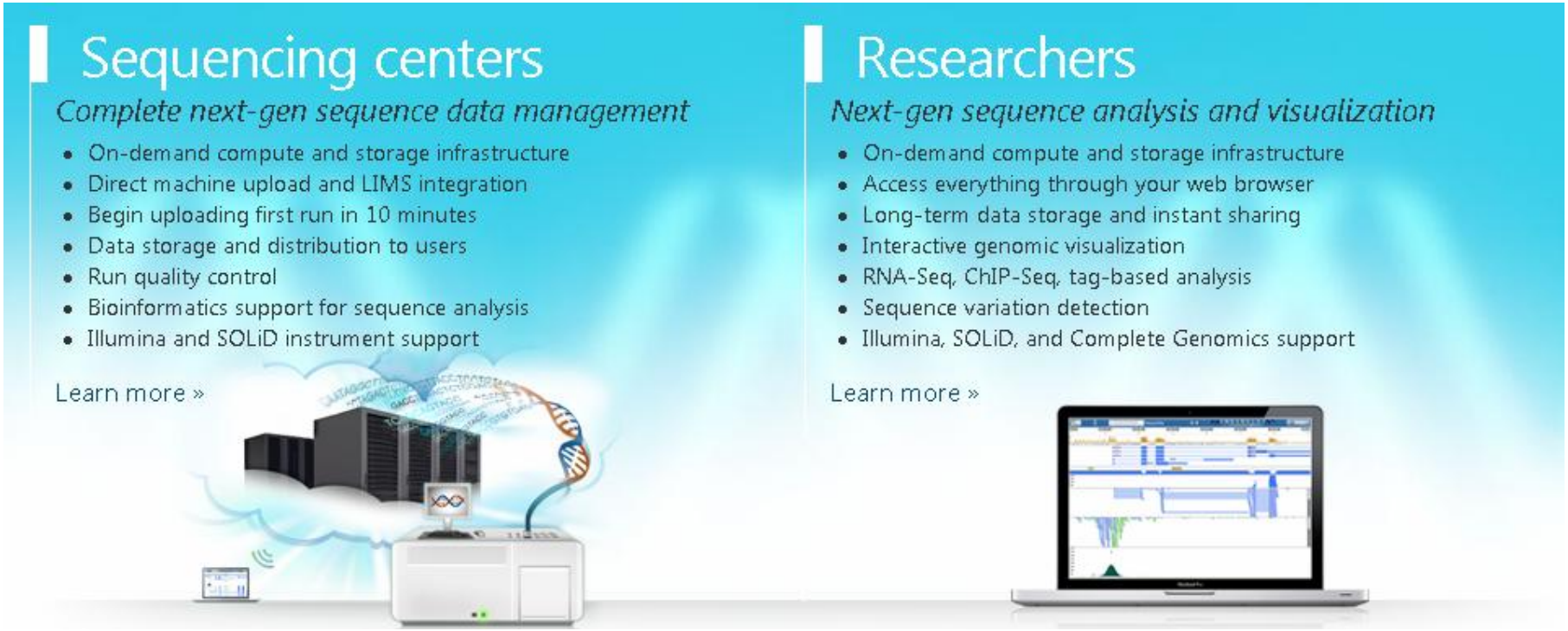
DNAnexus

Research lab

Research lab

Research lab

Features



- 不需軟體，上網即能分析與觀看結果
- 無電腦規格限制，雲端運算平行處理大量資料
- 使core facility容易發佈數據，減低電腦成本、維護成本與分析負擔。
- 使生物研究者能方便快速的自行整理NGS數據。
- 依數據量一次計費，1年內使用者自行多次免費運算。
- 支援fastq, csfasta, BAM, SAM格式。

NGS 序列數據簡介

Re-sequencing 數據分析流程簡圖

Start

Zipped 5GB → Zipped 3GB → 1 MB

End

原始序列

GATCATGTCACTATAC
G

Zipped 5GB

對應到 Ref Genome 上 (Mapping)

Zipped 3GB

計算區域內的短序列數目
分析mismatch位置與數量

以基因or片段為單位分析
改變差異, 以及交互作用

1 MB

RNA-seq Analysis Pipeline Comparison

8CPU+48GB RAM
工作站

Solexa SE, PE data
Fastq file

GATCATGTCACTATACG

SOLiD SE, PE data
csfasta + QV file

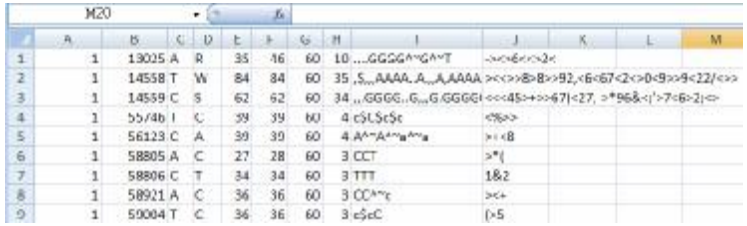
筆記型電腦

Mapping & Finding splicing
Tophat + Bowtie
BAM or SAM file

Quantify expression level
Cufflinks
Exp, Gtf files

SAMtools
Mutation Data

Txt Output



Txt Output

DNAnexus

RUN QUALITY METRICS
READ MAPPING
ANALYSIS RESULTS

Easy Access
Graphic Result

To Map or Not to Map…

Raw Reads (Before mapping) Raw Sequence Data

- Fastq
- Csfasta + QV

Mapped Reads (After mapping) Mapped (localized) Sequence

- SAM
- BAM

Fastq

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
3 +
4 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description

Line 2 is the *raw sequence* letters

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier

Line 4 encodes the quality values for the sequence in Line 2

Illumina sequence identifiers

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

csfasta +QV / csfastq

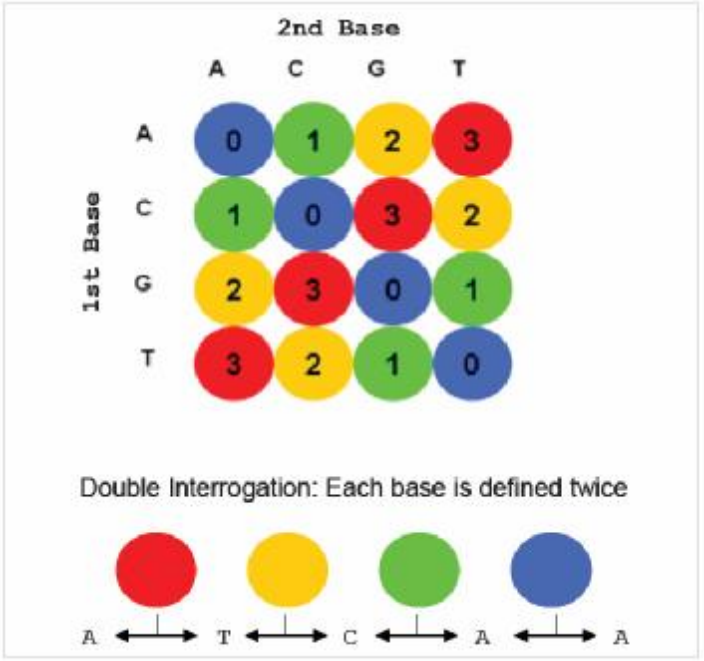
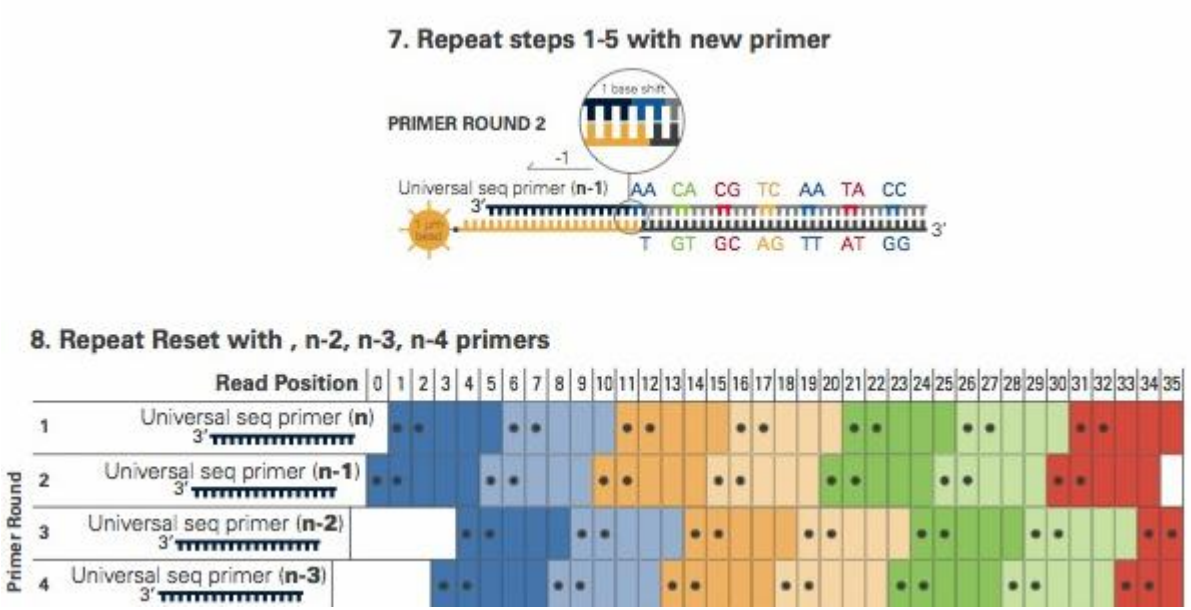


Figure 1: SOLiD Color Space Code



csfastq

@ERR000451.1 VAB_S0103_20080915_542_14_17_70_F3
330232302031021032233300203002330 01
%245719<.6353&:%0#$1%&%2(--27*%&%,

@ERR000451.2 VAB_S0103_20080915_542_14_17_171_F3
233203321200020012022100000000200 01
#&#$##&#&%%$#&#%##'#&$#%$*&-))##%')

Sequence Alignment/Map (SAM)

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

Header section:
Each header line begins with character @ .

HD – header
SQ-Sequence dictionary
RG-read group
PG-Program

The alignment section consists of multiple TAB-delimited lines with each line describing an alignment. Each line is:

```
<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL> \
  [<TAG>:<VTYPE>:<VALUE> [...]]
```

http://samtools.sourceforge.net/SAM1.pdf

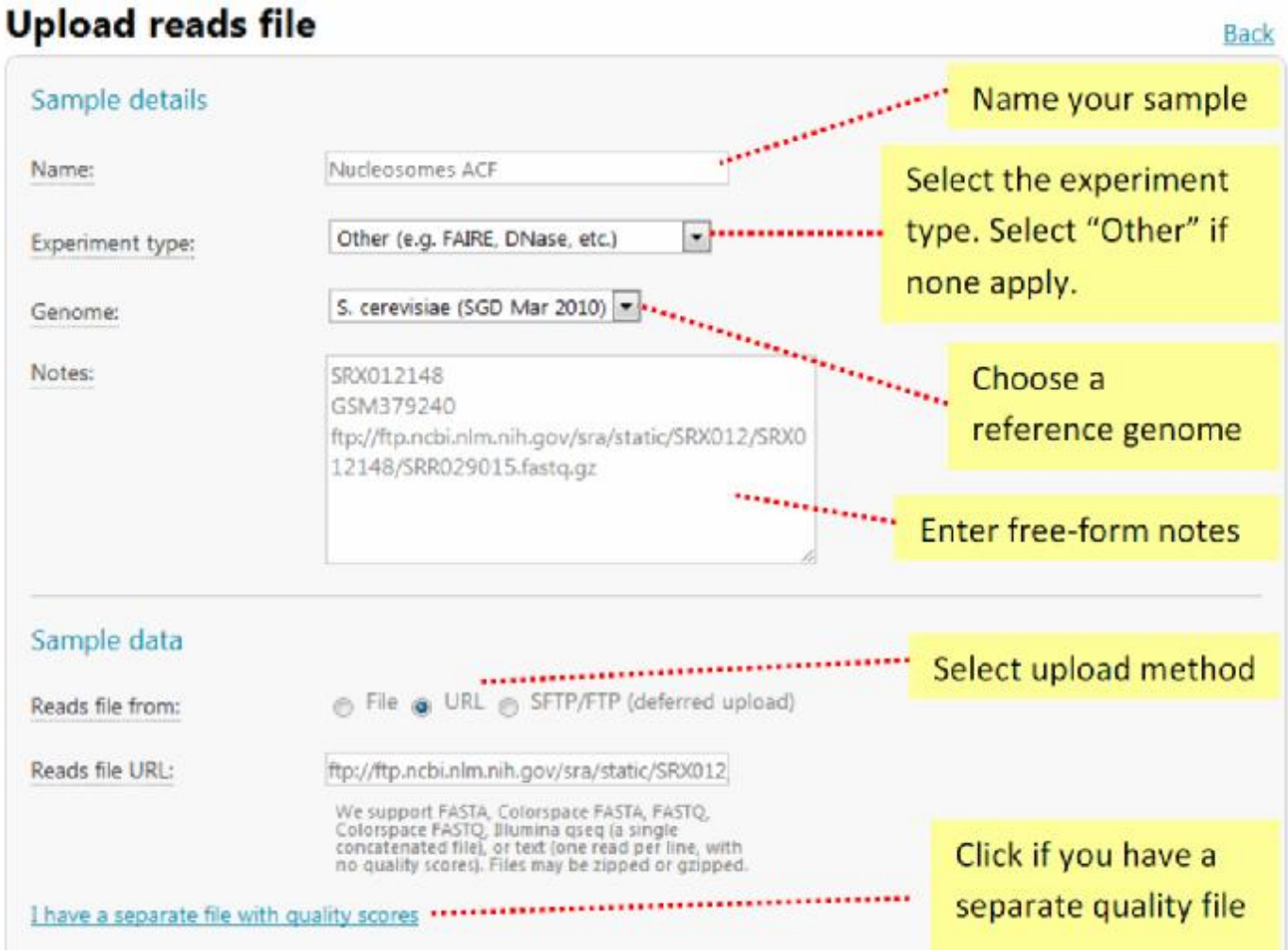
BAM

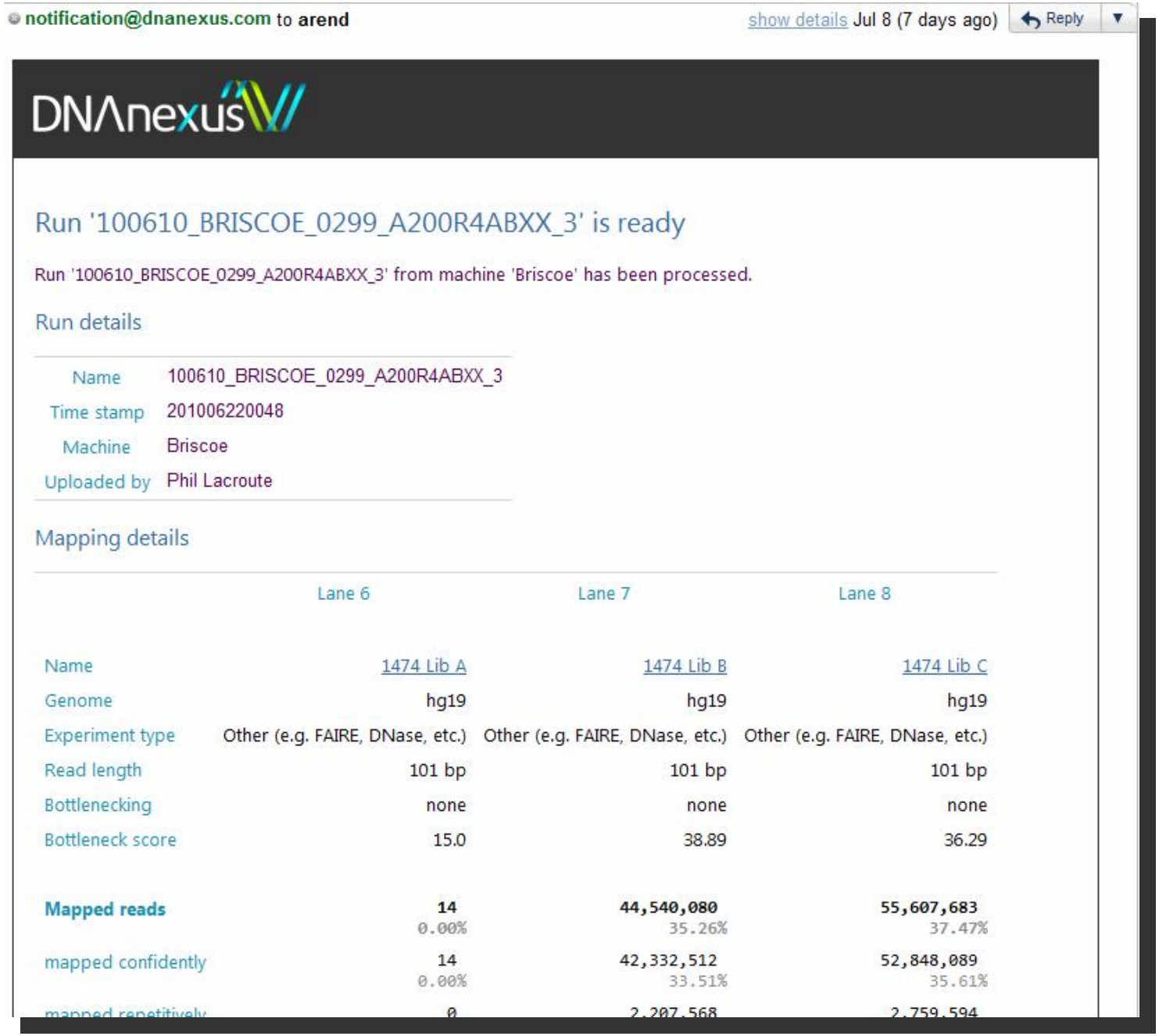
- BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
- BAM is compressed in the BGZF format
- The goal of BGZF is to provide good compression while allowing efficient random access to the BAM file for indexed queries.

DNAnexus操作流程

1. 登入帳號
2. 上傳數據
3. 點選數據進行分析
4. 點完侯即可離線，不需上線等候運算

Web Browser Upload
FTP/SFTP Upload





notification@dnanexus.com to arend

show details Jul 8 (7 days ago) Reply

DNAnexus

Run '100610_BRISCOE_0299_A200R4ABXX_3' is ready

Run '100610_BRISCOE_0299_A200R4ABXX_3' from machine 'Briscoe' has been processed.

Run details

Name	100610_BRISCOE_0299_A200R4ABXX_3
Time stamp	201006220048
Machine	Briscoe
Uploaded by	Phil Lacroute

Mapping details

	Lane 6	Lane 7	Lane 8
Name	1474 Lib A	1474 Lib B	1474 Lib C
Genome	hg19	hg19	hg19
Experiment type	Other (e.g. FAIRE, DNase, etc.)	Other (e.g. FAIRE, DNase, etc.)	Other (e.g. FAIRE, DNase, etc.)
Read length	101 bp	101 bp	101 bp
Bottlenecking	none	none	none
Bottleneck score	15.0	38.89	36.29
Mapped reads	**14** 0.00%	**44,540,080** 35.26%	**55,607,683** 37.47%
mapped confidently	14 0.00%	42,332,512 33.51%	52,848,089 35.61%
mapped repetitively	0	2,207,568	2,759,594

Quality control

Was my run good?

If not… why?

- Sufficient starting DNA
- rRNA contamination
- Base call quality distribution
- Paired-end library quality
- Coverage uniformity

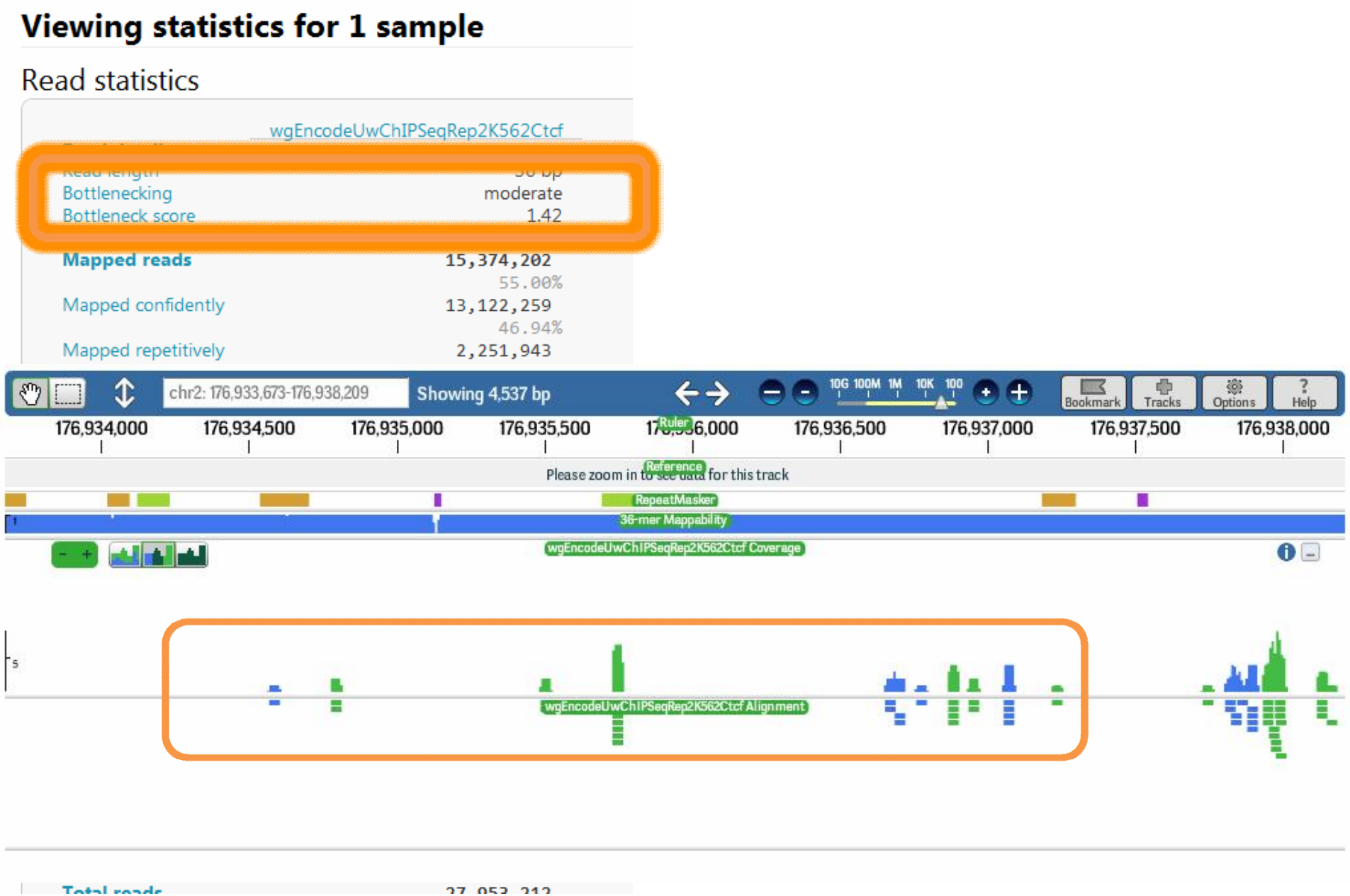
Quality control

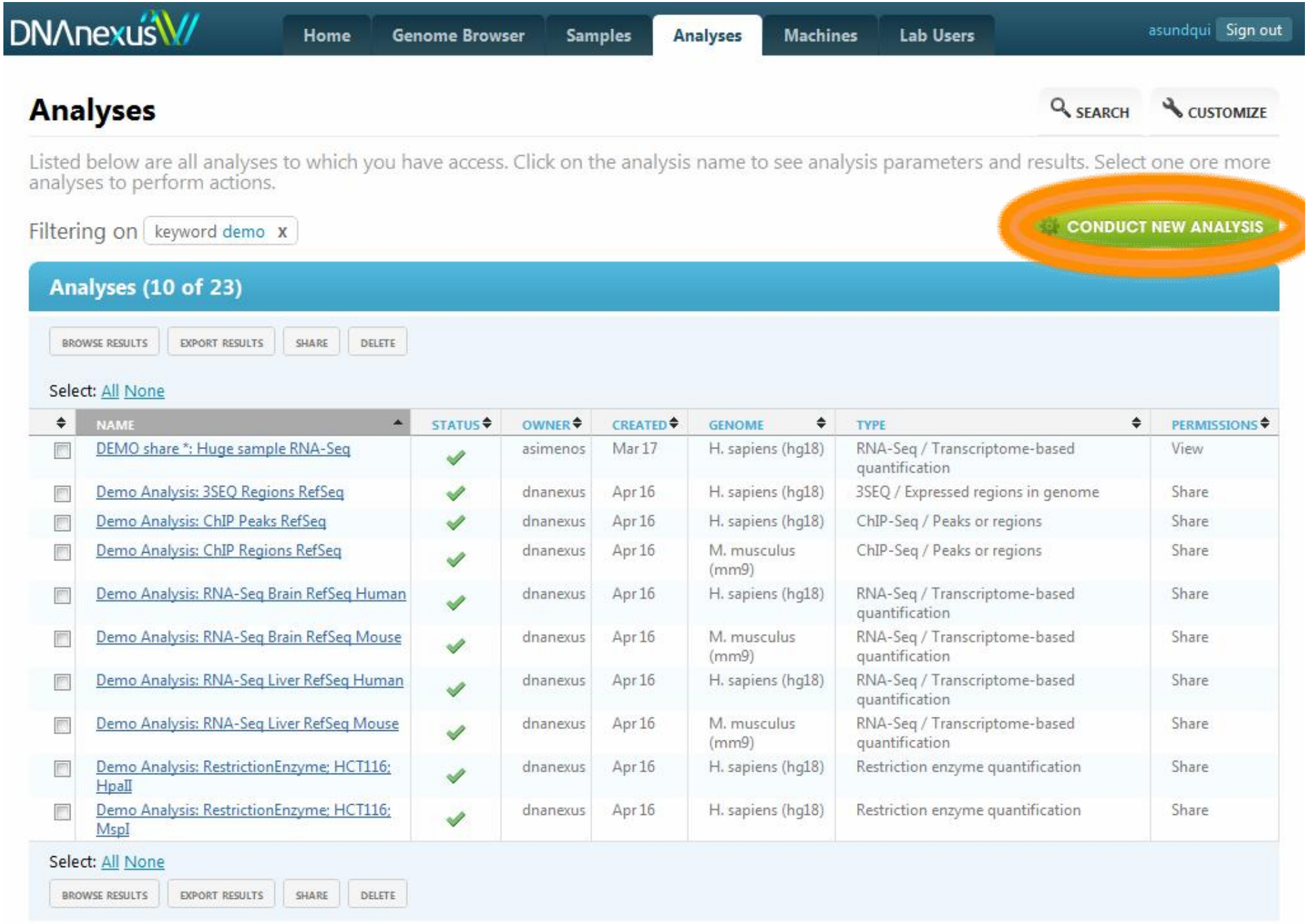
Viewing statistics for 4 samples

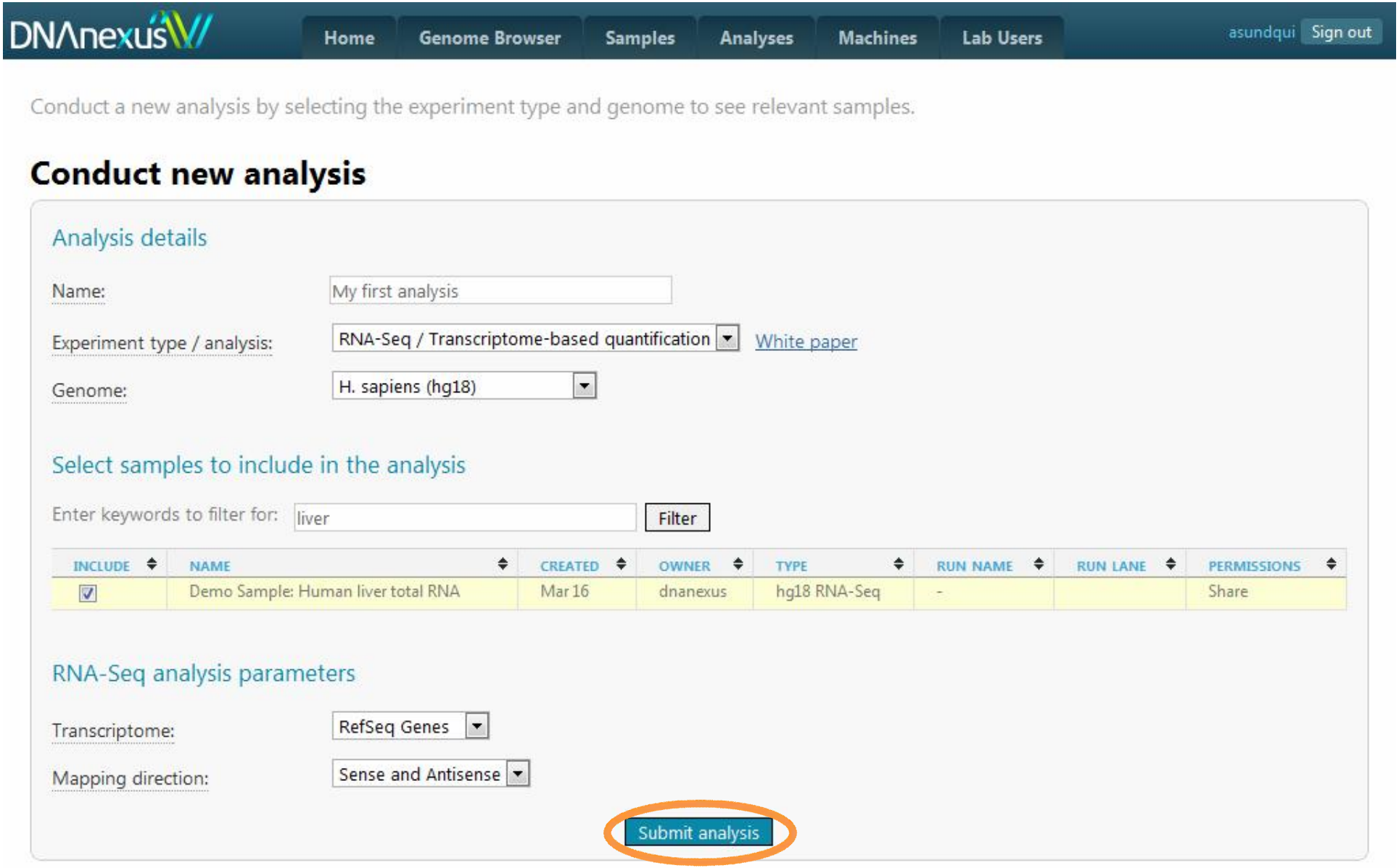
Read statistics

	AA3A-RNA-200bp-R1-24bp	AA3A-RNA-200bp-R2-24bp	AA3B-gDNA-500bp-HS-R1	AA3B-gDNA-500bp-HS-R2
Read details				
Read length	24 bp	24 bp	101 bp	101 bp
Bottlenecking	none	none	none	none
Bottleneck score	54.91	53.32	5.07	5.34
Mapped reads	**1,720,473**	**1,741,597**	**34,184,135**	**28,655,593**
	9.51%	9.63%	63.27%	53.03%
Mapped confidently	925,371	936,460	29,691,554	23,943,881
	5.12%	5.18%	54.95%	44.31%
Mapped repetitively	795,102	805,137	4,492,581	4,711,712
	4.40%	4.45%	8.31%	8.72%
Reads not mapped	**16,366,784**	**16,345,660**	**19,848,064**	**25,376,606**
	90.49%	90.37%	36.73%	46.97%
no mapping	536,818	576,037	12,240,651	19,093,717
	2.97%	3.18%	22.65%	35.34%
low quality	56,403	372,009	7,292,647	5,966,225
	[illegible]	[illegible]	[illegible]	[illegible]
ribosomal RNA	15,767,755	15,392,529	307,188	287,866
	87.18%	85.10%	0.57%	0.53%
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0.02%	0.01%	0.01%	0.05%
control (phiX-174)	0	0	0	0
	0.00%	0.00%	0.00%	0.00%
poly-A	699	1,550	342	885
	0.00%	0.01%	0.00%	0.00%
poly-C	294	975	92	73
	0.00%	0.01%	0.00%	0.00%
poly-G	196	224	0	0
	0.00%	0.00%	0.00%	0.00%
poly-T	239	430	32	30

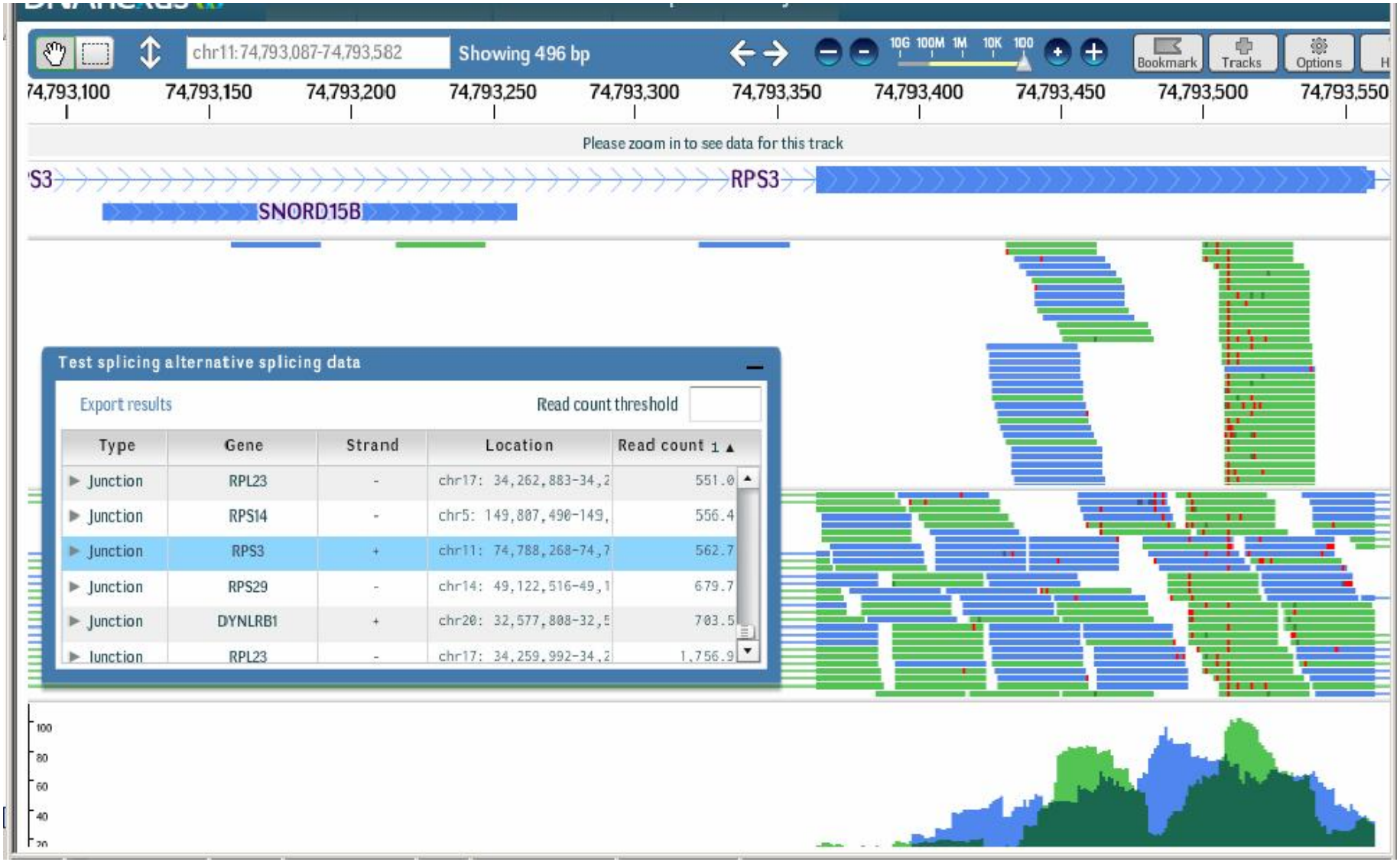
Quality control

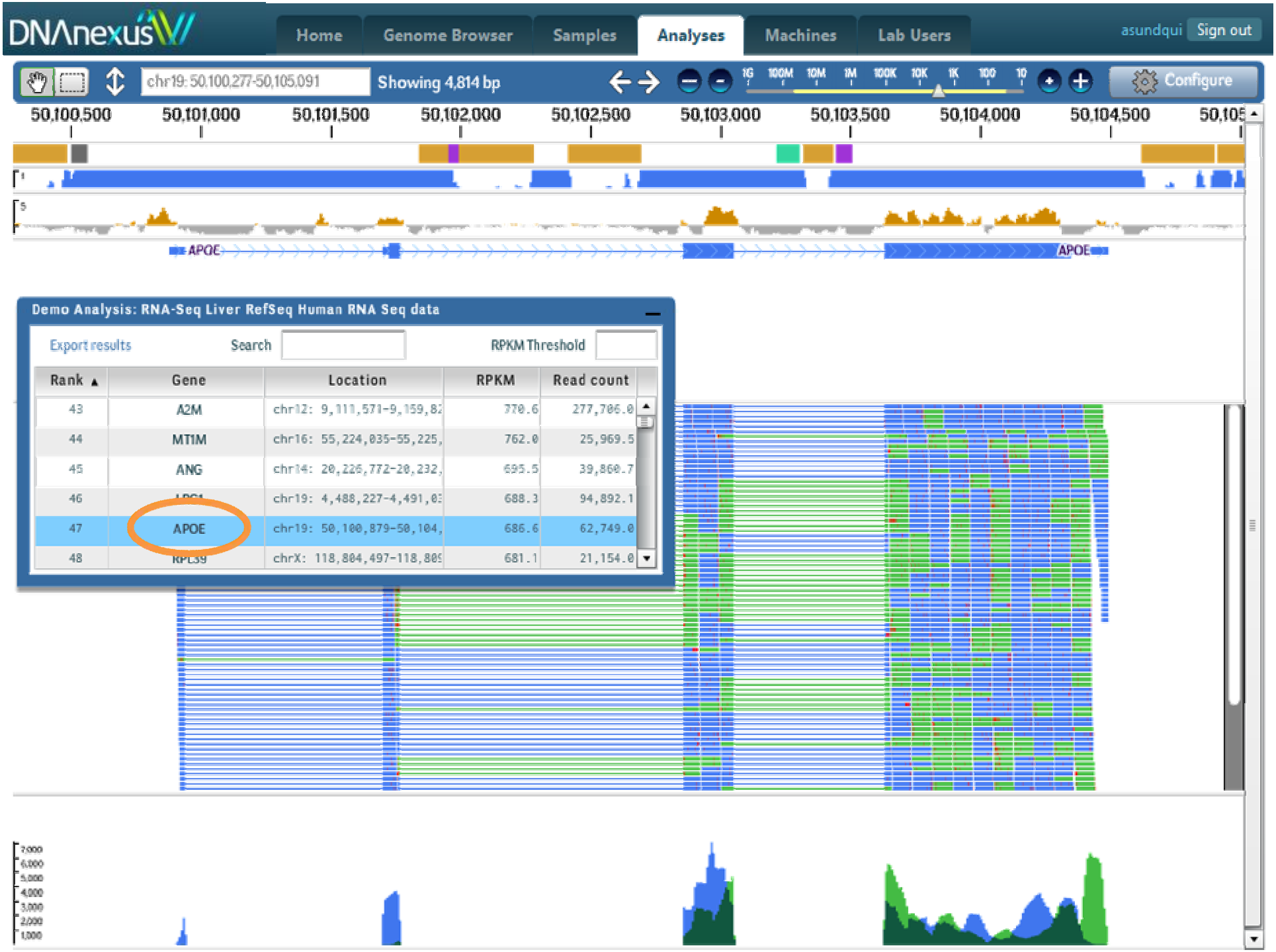




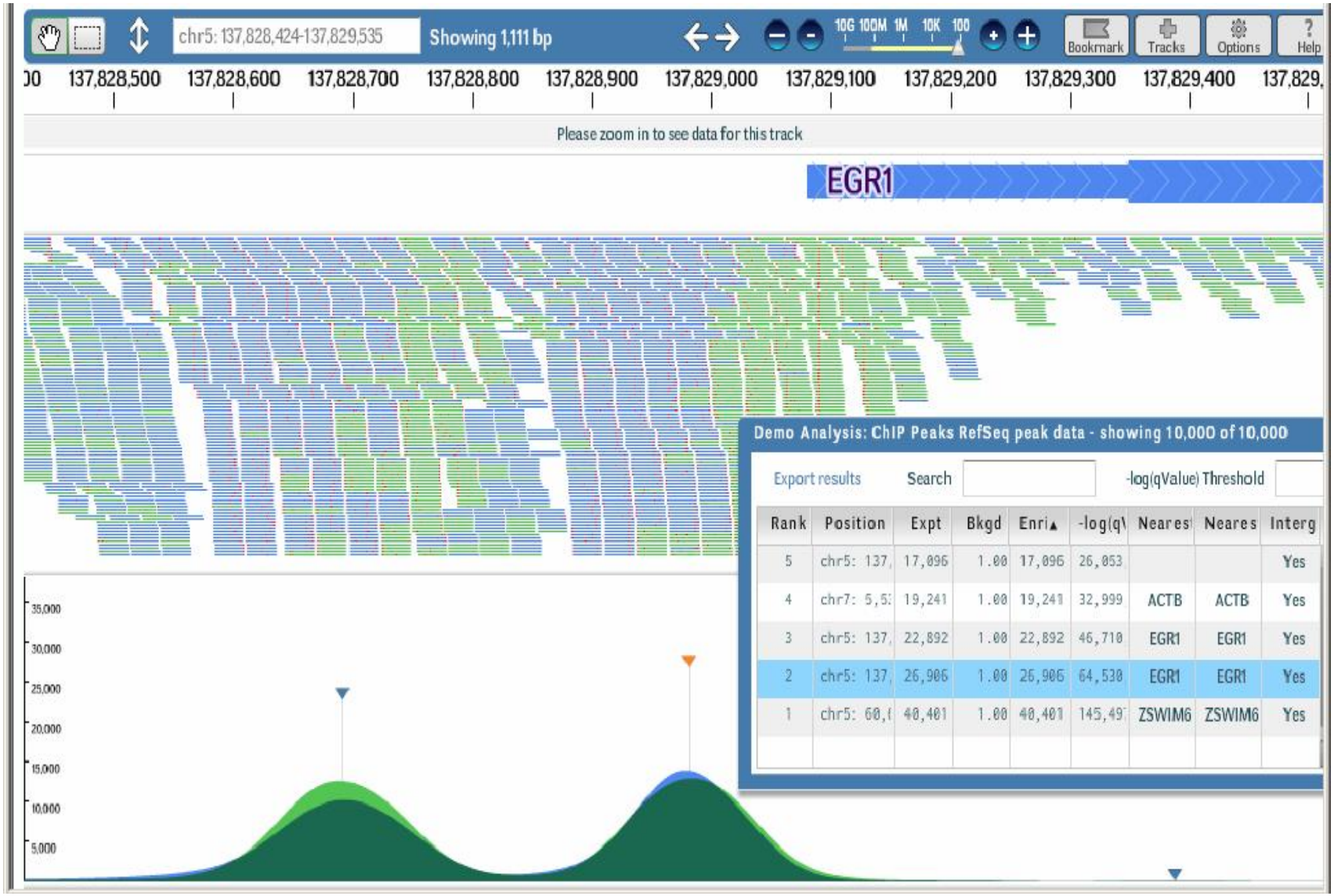


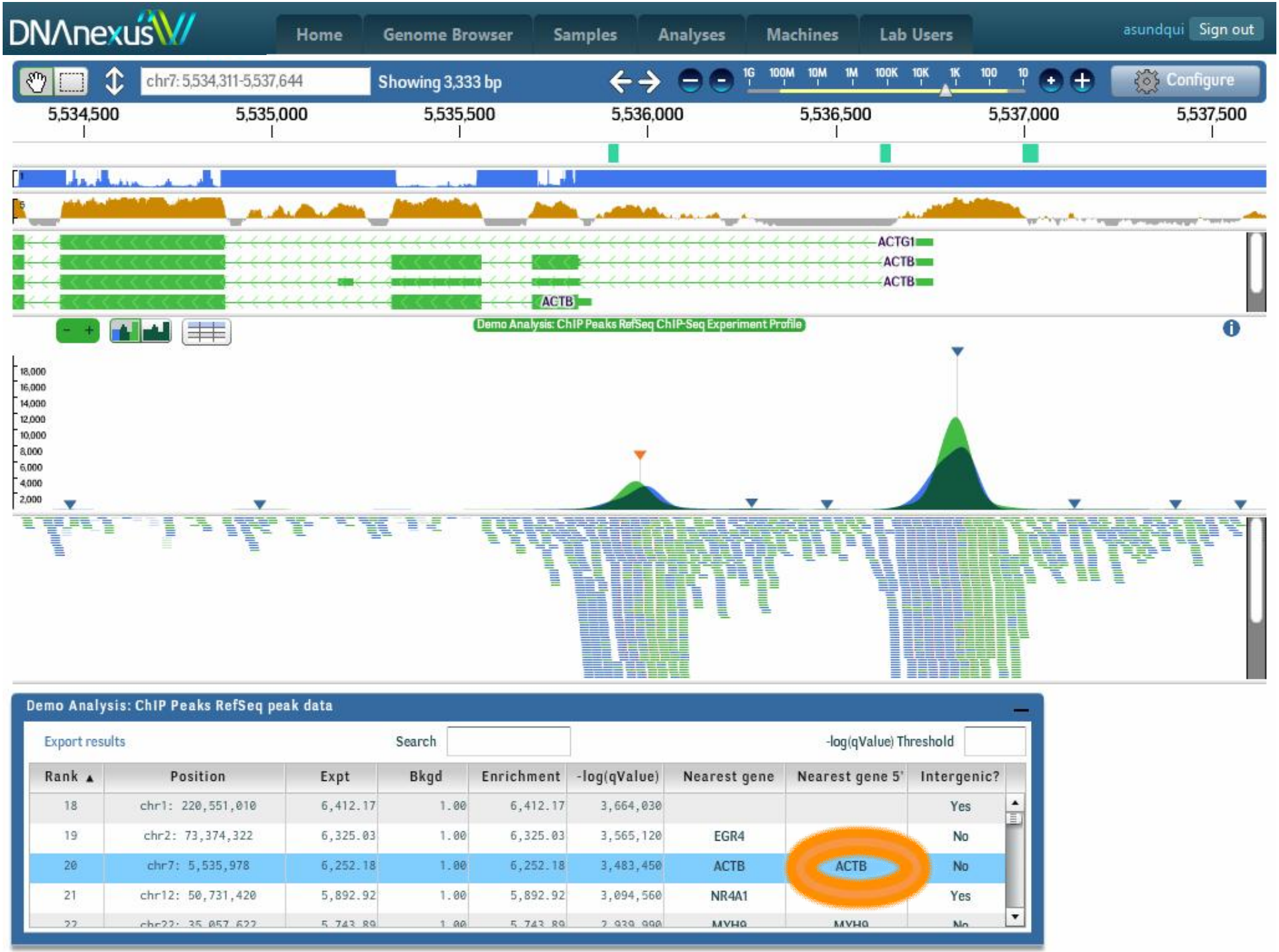
RNA-seq Analysis



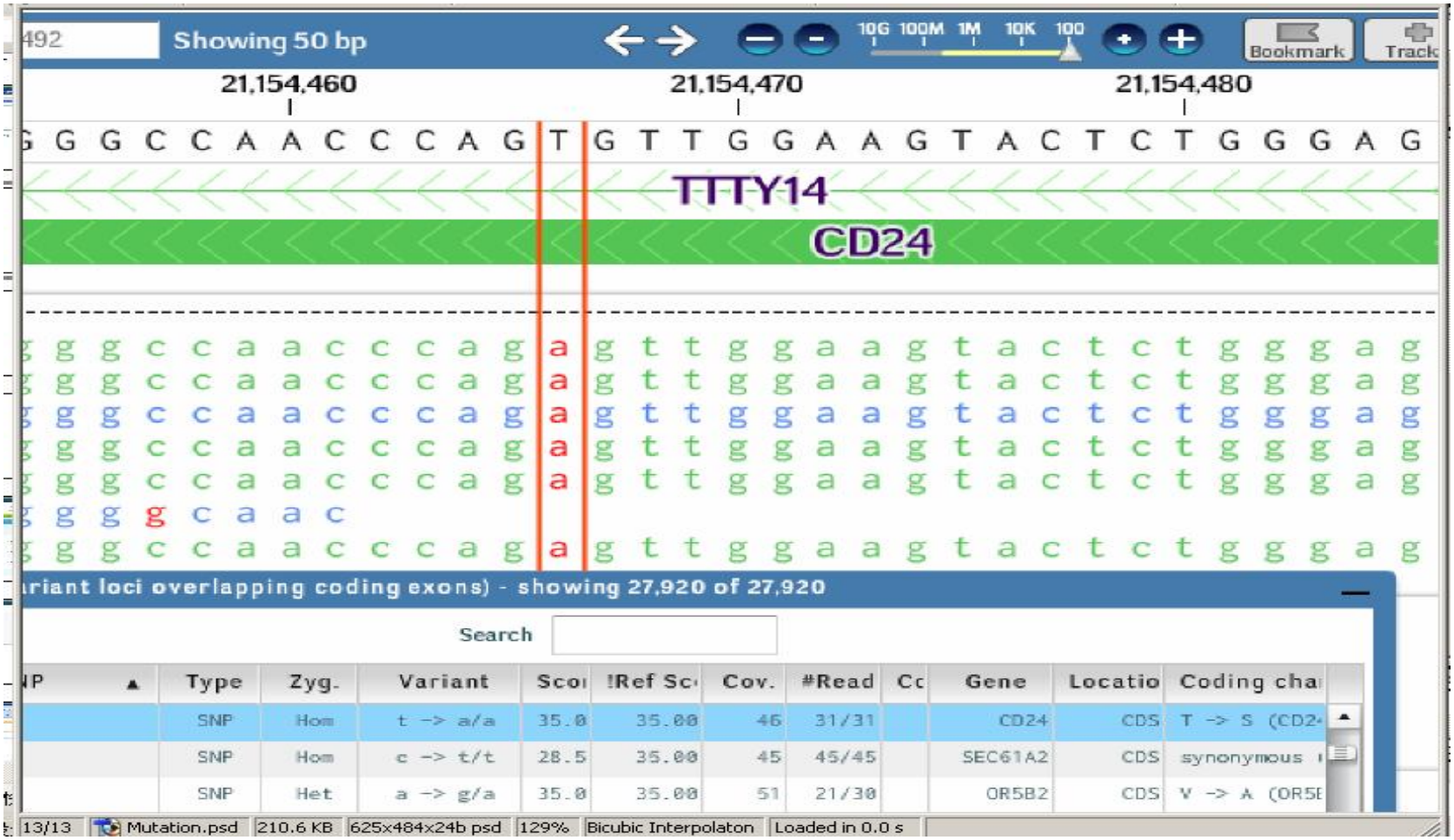


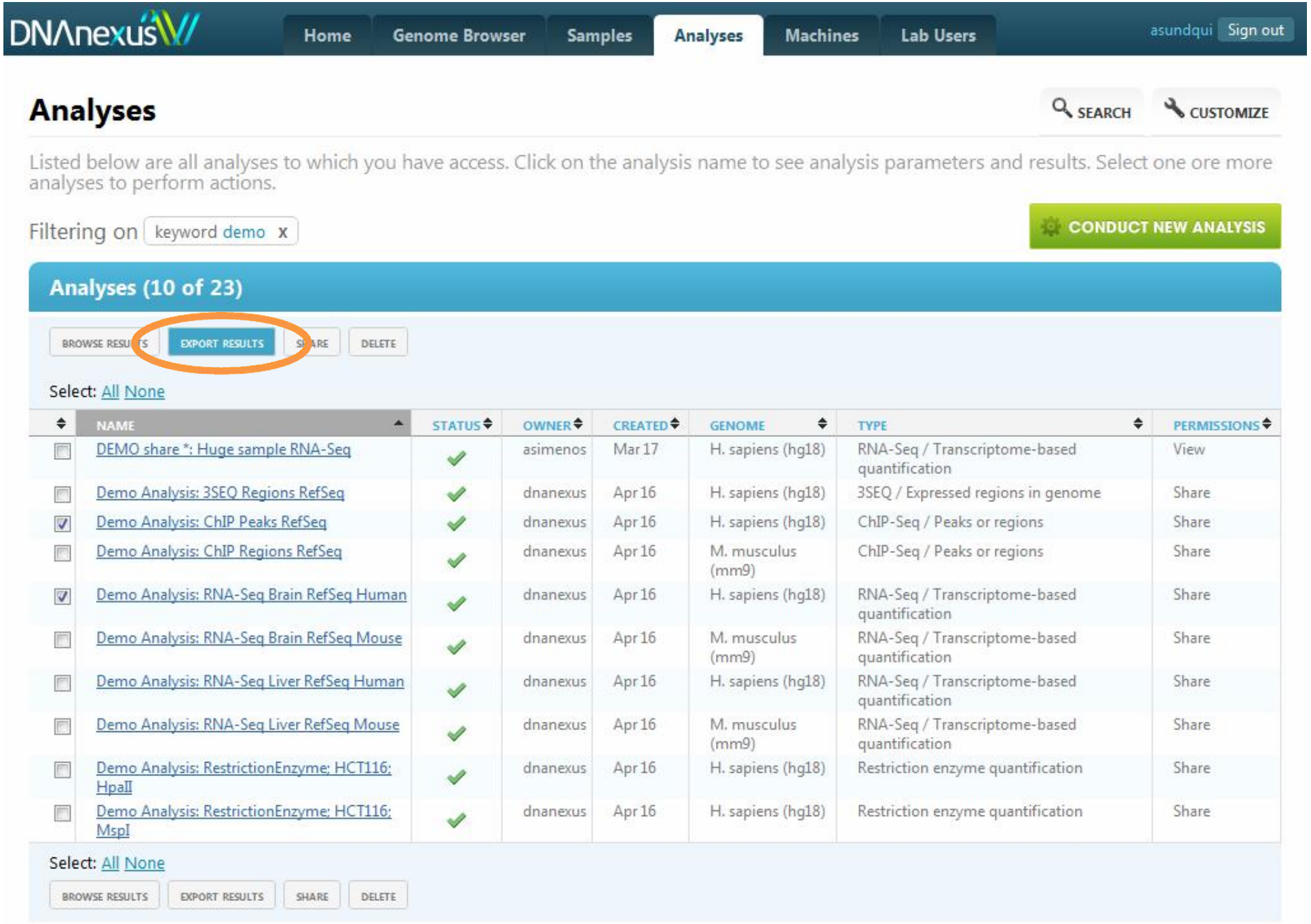
ChIP-seq Analysis

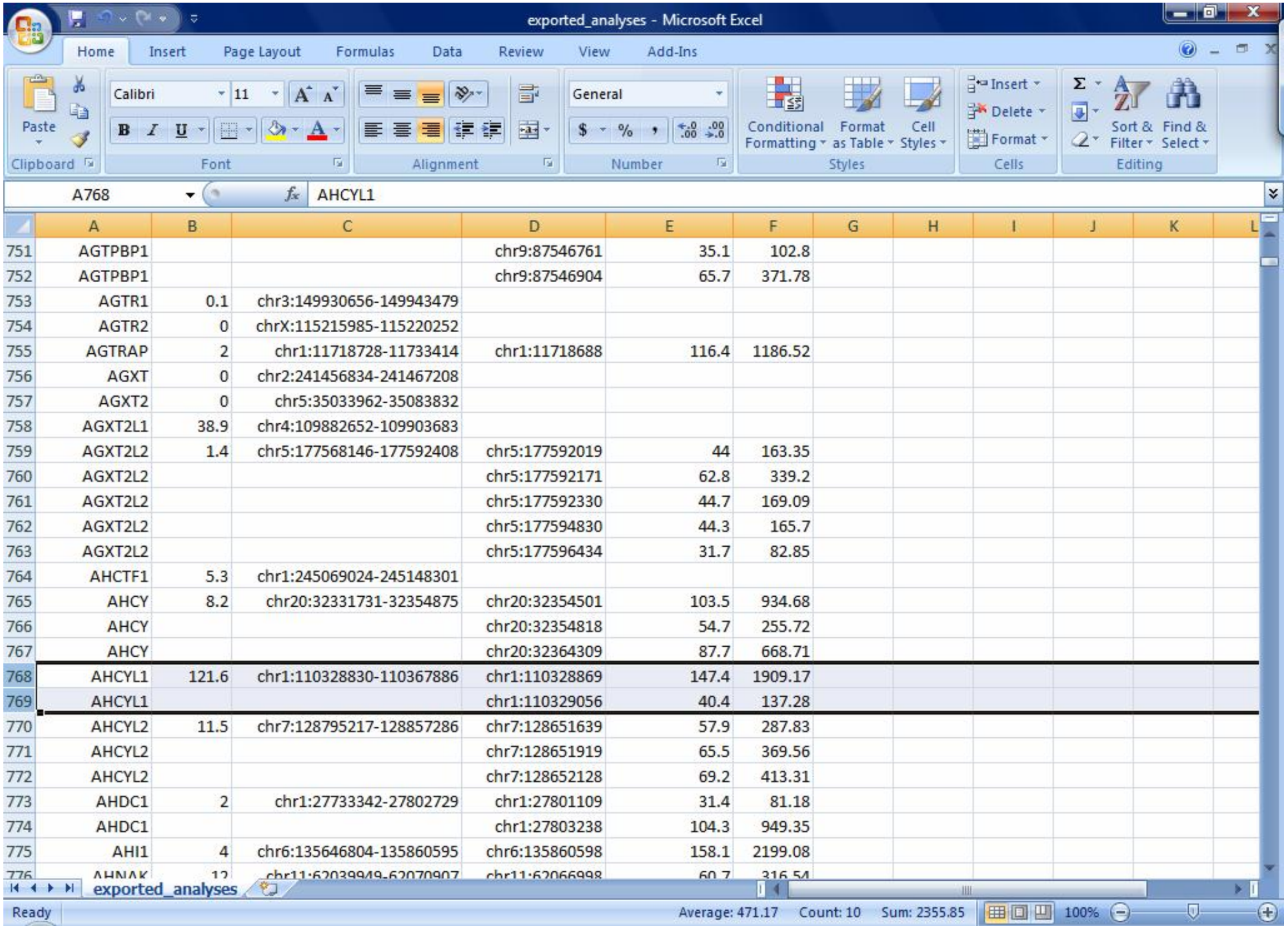




Mutation Analysis





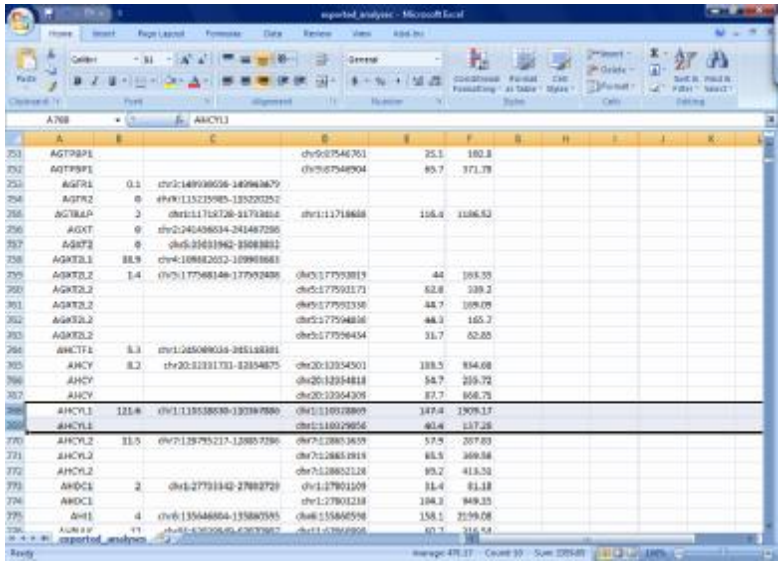


	A	B	C	D	E	F	G	H	I	J	K	L
751	AGTPBP1			chr9:87546761	35.1	102.8						
752	AGTPBP1			chr9:87546904	65.7	371.78						
753	AGTR1	0.1	chr3:149930656-149943479									
754	AGTR2	0	chrX:115215985-115220252									
755	AGTRAP	2	chr1:11718728-11733414	chr1:11718688	116.4	1186.52						
756	AGXT	0	chr2:241456834-241467208									
757	AGXT2	0	chr5:35033962-35083832									
758	AGXT2L1	38.9	chr4:109882652-109903683									
759	AGXT2L2	1.4	chr5:177568146-177592408	chr5:177592019	44	163.35						
760	AGXT2L2			chr5:177592171	62.8	339.2						
761	AGXT2L2			chr5:177592330	44.7	169.09						
762	AGXT2L2			chr5:177594830	44.3	165.7						
763	AGXT2L2			chr5:177596434	31.7	82.85						
764	AHCTF1	5.3	chr1:245069024-245148301									
765	AHCY	8.2	chr20:32331731-32354875	chr20:32354501	103.5	934.68						
766	AHCY			chr20:32354818	54.7	255.72						
767	AHCY			chr20:32364309	87.7	668.71						
768	AHCYL1	121.6	chr1:110328830-110367886	chr1:110328869	147.4	1909.17						
769	AHCYL1			chr1:110329056	40.4	137.28						
770	AHCYL2	11.5	chr7:128795217-128857286	chr7:128651639	57.9	287.83						
771	AHCYL2			chr7:128651919	65.5	369.56						
772	AHCYL2			chr7:128652128	69.2	413.31						
773	AHDC1	2	chr1:27733342-27802729	chr1:27801109	31.4	81.18						
774	AHDC1			chr1:27803238	104.3	949.35						
775	AHI1	4	chr6:135646804-135860595	chr6:135860598	158.1	2199.08						
776	AHNAK	12	chr11:62039949-62070907	chr11:62066998	60.7	316.54						

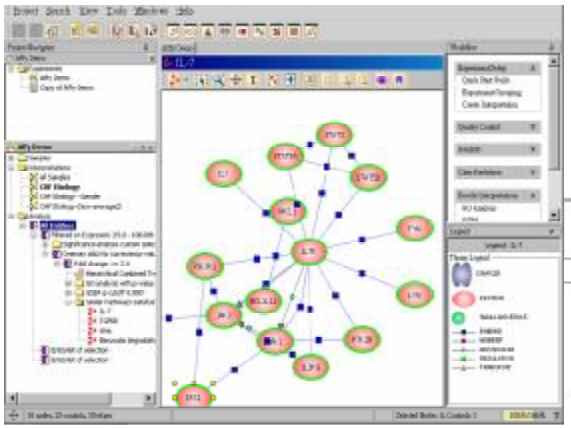
後續分析搭配

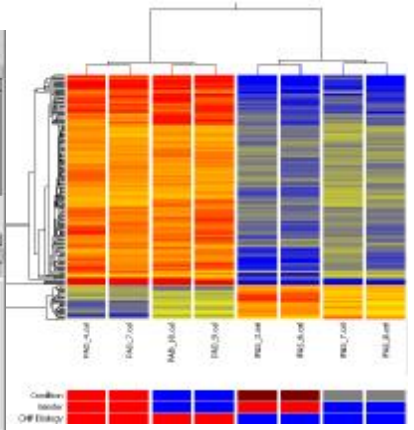
DNAnexus

Excel



GeneSpring





Thank you for attending !!

Wish you have a pleasant research~

Mapping

The read mapping method is similar to other pattern-based read mappers, including ELAND, ZOOM, and MAQ.

Heuristic approaches such as k-mer counting and seed-based algorithms have been shown to work similarly well with greatly reduced computational cost

As the best quality scores typically occur in the first cycles of a sequencing run, our pattern matching focuses on the base calls in the first 36 bases (or up to the read length if it is shorter). Thus, we guarantee mappings of all reads to all genomic locations with 0, 1, or 2 mismatches in the first 36 bases of the read. Additional mismatches may occur either in this seed region or in the latter part of the read.

method	reads/s	% "best mapping" incorrect				% true source not found			
		0 SNPs	1 SNP	2 SNPs	3 SNPs	0 SNPs	1 SNP	2 SNPs	3 SNPs
DURA	4131	**3.10**	**4.06**	**6.30**	**14.49**	**0.06**	**0.34**	**1.89**	**5.80**
Bowtie (speed)	**40690**	4.37	8.86	21.03	44.93	4.37	8.86	21.03	44.93
Bowtie (accuracy)	4605	3.33	5.44	9.93	23.19	0.82	2.59	7.15	15.94
BWA (speed)	8421	3.19	4.56	8.66	23.19	0.63	1.36	5.18	17.39
BWA (accuracy)	6451	3.15	4.13	6.53	**14.49**	0.21	0.57	2.40	8.70
MAQ	669	3.47	6.18	16.27	42.03	3.47	6.18	16.27	42.03

Table 1. Benchmarks of read mapping accuracy and computational efficiency in simulated whole genome human resequencing. We benchmark DURA against Bowtie, BWA, and MAQ. Best results are shown in bold.

source genome	method	% unique mapping	% 2 mappings	% 3-10 mappings	% >10 mappings	% no mapping	% miss mapping
hg18	DURA	**94.96**	1.38	1.38	2.27	**0.020**	**0.03**
	Bowtie	92.64	2.70	1.93	2.60	0.131	0.83
	BWA	94.56	1.41	1.52	2.48	0.025	0.18
hg18evo	DURA	**94.29**	1.39	1.39	2.29	0.639	**0.85**
	Bowtie	92.44	2.69	1.92	2.52	0.434	1.75
	BWA	94.25	1.42	1.54	2.52	**0.276**	0.99

3SEQ/ RNA-seq

Once the reads have been mapped to the transcripts, each transcript is quantified by calculating its RPKM value (reads per kilobase of transcript per million mapped reads; Mortazavi et al., 2008). RPKM is defined as follows: If the number of reads that map to a given transcript t is Mt, the length of that transcript is Lt, and the total number of mapped reads is M, such that $M = \sum Mt$, then $RPKM = (109 * Mt)/(Lt*M)$.

The 3SEQ / transcriptome analysis is a variant that focuses on quantification of transcripts in libraries produced with the 3SEQ protocol (Beck et al., 2010). 3SEQ libraries are constructed such that there is one read per transcript, which originates near the 3' end usually in the 3' UTR. Reads produced from these libraries will concentrate on the annotated 3' UTRs when mapped to the transcriptome (and do not typically span the whole gene like in an RNA-Seq analysis). Because there is one read per transcript molecule, calculating RPKM values is inappropriate and only the read counts (weighed by the posterior probability of their mapping) are reported for each gene. Normalization by the number of reads in the sample, or by calculating a Z score, should be performed on the reported read counts before comparisons among samples. For genes with more than one transcript, the transcript with the highest read count is chosen to represent the gene.

ChIP-seq

Similar to the QuEST method, DNAnexus uses kernel density estimators (KDEs) to integrate closely spaced read mappings. we use only confidently mapped reads with posterior probability greater than 90% to compute the density. The breadth of the kernel's distribution can be adjusted by the kernel bandwidth parameter; larger values cause a greater degree of smoothing of the density profile, leading to more contiguous regions. We typically recommend a kernel bandwidth of 30 for transcription factors, 60 for RNA polymerase II-like factors, and 100 for histones.

The DNAnexus ChIP-seq algorithm appropriately uses the background sample to estimate read enrichment over background, calculate statistical significance (as q--values), and estimate a false-discovery rate (FDR). The false discovery rate is then the ratio of these two: FDR = |Peaks(experiment=B1, background=B2)| / |Peaks(experiment=E, background=B2)|.

Nucleotide-Level Variation

- This is done considering the contents of the reads overlapping each position of the genome, and reporting the most likely differences in the sample's DNA that could lead to this sequencing result. Differences include single- and multi-nucleotide polymorphisms (called SNPs and MNPs, respectively), insertions, and deletions. For ease of nucleotide level data viewing, the results are annotated with specific coding changes in the genome, and include summary evolutionary statistics for the sample analyzed.
- DNAnexus' indel module can handle indels up to 10 bp

Type	Description	Some examples (Reference -> Genotype) and their meanings
SNP	A single nucleotide is mutated	A -> A/G (heterozygous SNP) A -> G/G (homozygous SNP) A -> G/T (heterozygous SNP, new alleles)
MNP	A series of consecutive nucleotides are mutated	AA -> AA/CT (heterozygous MNP) AA -> CT/CT (homozygous MNP) AA -> AT/CA (heterozygous MNP, different phase than AA/CT)
INS	Sequence in inserted	-> A/A (homozygous insertion) -> /A (heterozygous insertion) A -> AA/AA (homozygous insertion, exact location is ambiguous)
DEL	Sequence is deleted	A -> / (homozygous deletion) A -> A/ (heterozygous deletion) AA -> A/A (homozygous deletion, exact location is ambiguous)
MIXED	Any other combination of events	A -> /C (deletion and mutation) ACA -> T/T A -> CC/CC

Population Allele Frequency Analysis

- DNAnexus now provides Population Allele Frequency analysis. This analysis can be performed on groups of one or more samples. Each group represents a population, and the output includes variant allele frequencies across populations. The data reported in the output lists the location and frequency of all variants identified. For each genomic location with variation, the two most frequent alleles X and Y across all populations are identified, and the frequencies of the three possible genotypes (X/X, X/Y, and Y/Y) are summarized for each population. Listed in separate columns for each group are the frequencies for "other" (number of group members whose genotypes are not X/X, X/Y or Y/Y) and "unknown" (number of group members for which there was no variation call due to insufficient coverage). The results also contain gene annotations, and a P-value of a chi-square test indicating whether allele frequency distributions differ among groups.

Exome Analysis

- The newly added Exome analysis computes key coverage statistics for each exon in a set of genomic regions defining an exome. For this analysis, both vendor-supplied (Agilent and Nimblegen) and custom user-uploaded exomes are supported. User-supplied exomes must be provided in BED file format. For each exon, the number and fraction of bases covered by sequence reads are reported, along with the average coverage within the exon. Exons overlapping genes in a gene annotation track are labeled with the gene name to allow easy searching for exons from a gene of interest.