



NGS 雲端分析簡介

李彦樑 (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^{*}



Workflow & Specs

HiSeq 2000

The HiSeq 2000 seque and a breakthrough us proven and widely-add sequencing by synthes engineering, HiSeg 201 sequencing output and interaction design feat set a new standard for



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
	Read Length 1 × 35 bp 2 × 50 bp 2 × 100 bp	Read Length Run Time 1 x 35 bp ~1.5 days 2 x 50 bp ~4 days 2 x 100 bp ~8 days

From www.Illumina.com

Library Type	Read Length	Days/Run	Total Tags/Run	Mappable Data
	2 x 35 bp	8-9	» 1,4 B	50-70 GB
Mate-Paired	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
E CONTRACTOR CONTRACTOR	1 x 35 bp	3.5-4.5	⇒700 M	25-35 GB
Fragment	1 x 50 bp	6-8	>700 M	40-50 GB

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 \times Read Length = No. of Base 75.7 M reads \times 50 base = 3.79G base



Sequencing is only the beginning



Compute resources?



Sequence Analysis "Gap"







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





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
- Hpall/Mspl 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

											And in sec.	CCTCOTOR.	
										SLOGOT	and a second	CIC COLOR	and the second s
										30-100	ORCE	New Constant	and Pa
						1.57%			_		10 TOI		6
ang enelysis 'Zebrahs	h FIP fest - 1	Findows Internet	splorer		and the last								6
 Manane 	ous.completings	stosharu mayar	arkae"igewoe	Careford monorphy	ALL INTERO			-					1
ERWIE (COR(Y)	RUDWR TO (A)	1.14(1) 18.99(9)							-			00	
ilitile 🙀 🖅 Skydmo								1			-		
Analyses		Browsing analysis 'Ze	oraf 🛪 👔	Browsing analy	sis 'Zebrafis	h	→ ★主任(5)		-	-	100	O MARK	arran a
nexus W/	Home	Bookmarks	Seno mes	Samples	Analyse	22	jackiezsia S	1	JØ.				
	A CONTRACTOR OF		Utonatuli	0000000	4								
Chr1:27,	76,458-27,076,	80 Showi	ng 112 bp		\leftrightarrow			an Phel					
0 27,076,480	27,076,490	27,076,500	27,076,510	27,076,520	27,07	6,530 27,076,540 27,076,550 27,07	76,560 27,076	570 2	97,071				
I TODAT TOTOO	I	1	I	1	OTTON		I I						
A. DURAPHI INTERACTORY	JOHON I BCC	300110A0ACA0	01001000	ALLATEIGET	CLUCK	GIACITINGGUICATAAATATGCAGCTTGG	CICCI ICUI CCA	U CANA I	12/01				
	a contraction	*****		qdh	「日本市社会			igdh)	10.00				
		******		egdh	SEE			igdh) i					
				igdh)			(iqdh)					
				egdh)				iqdh) -					
				ngdh				ogdh)					
				ngdh 	G			vçdh)	ī				
		•		ngdh H	Git			vgdh)	I				
		•		egdh	Ce			sgih)	1				
Drafish FTP Test (Vari	int loci overs	spping cading exa	18) (7773 rov	egith	C +			sgdh	1				
Drafish FTP Test (Var)	int loci overl	ipping ceding exe	16) (7773 rov	rigith :	G P			vydh)					
prafish FTP Test (Vari part results asition Neorest Gene	nt loci overt Type	spping coding exo Zyg. Gase Relation	ns) 17773 rov Variant	vs) #Reads	C t	Coding Changes	18af Score	cov.	Conv	ſ			
orafish FTP Test (Var) uper results asition Neorest Gene r1: 85,5 doubld2	int loci overf Type SNF	zyg. Game Relation Het CE	(s) 7773 rov Variant g → s/g	vs) #Reads 5/4	Scare 5.52	Cading Changes zynony rous (dcun1dz/EMSDAFTE00000058971)	iRaf Score 25.00	cov.	Core				
orafish FIP.Test (Var) estition Neorest Gene r1: 85,5 dcunlo2 r1: 199, gati	int loci over l Type SVP SVP	Zyg. Gene Het CE Het CE	Variant g → a/g g → a/g	rgdh VS) #Reads 5/4 2/2	C 5.57 18.65	Coding Changes Tymosysous (dcun152/EXSD4FT000000669371) Tymosysous (gar6/D46C04T00000010012)	Illef Score 35.00 18.00	Cov.	Conv				
brafish FIP Test (Var) Sport results Pasition Nearest Gene vr1= 85,5 dcunld2 rr1= 199, gasti ur1= 217, agus[54001	int loci overl Type SNP SNP SNP	Zyg. Game Relation Hat CIT Hat CIT Hat CIT	Variant g → a/g g → a/g a ~ a/g	rgdh: vs) #Reads 5/4 2/2 3/18	C 5.52 18.65 27.69	Coding Changes Tyronysous [doub1d2/ENSDAPTee000058971] Tyronysous [doub1d2/ENSDAPTee000058971] Tyronysous (gas6/ENSCAPTee000018912) Tyronysous (gas1]14001/ENSDAPTe000000018127)	IBaf Score 25.00 14.00 25.00	Cov. 21 21	Core				
brafish FTP Test (Var) Synstressits Pasition Nearest Gene hr1: 85,5 deunloz hr1: 199, gasti hr1: 217, ags:154081	Type SNP SNP SNP SNP SNP SNP	Zyg. Game Relation Het CIS Hat CIS Hat CIS How CIS	Variant g -> a/g g -> a/g a -> a/g g -> c/c	rgdh: vs) #Reads 5/4 2/2 3/18 4/4	C 5.52 18.65 27.66 12.26	Coding Changes tyeonytous (doun1d2/ENSDATTeeeeeeeEPSTT) tyeonytous (gas6/ENEDATTeeeeeeEPSTT) tyeonytous (age:114401/DEEDATTeeeeeeE8227) A → G (sge:154001/ENSDATTeeeeee28227)	IBaf Score 25.00 13.00 25.00 25.00	Cov. 9 4 21 4	Conv				
Orafish FTP Test (Var) xpsrt results 'asition Nearest Gene r1: 85,5 dcunld2 r1: 199, gasti (r1: 217, agu;124001 r1: 217, agu;154001 r1: 724, agu;56685	Type SNP SNP SNP SNP SNP SNP SNP SNP	Zyg. Game Relation Het CI Hat CI Hat CI Hos CI Hos CI CI	Variant g -> a/g g -> a/g g -> c/c g -> a/g g -> c/c g -> a/a/b	rgdh: vs) #Reads 5/4 2/2 3/18 4/4 5/5	Scare 5.52 18.65 27.46 12.26 14.95	Coding Changes Tyronysous (doun!d2/ENSDAPTee000058971) tyronysous (gss6/DECAATe000001802) tyronysous (gss11:4001/DEDAAT0000003827) A → 6 (ssg: 1:54001/SESDAFT0000003827) tyronysous (zgc; 56655/ENSDAFT0000003827)	IBaf Score 25.00 43.00 25.00 25.00 25.00	Cov. 9 6 21 6	Core				
Byoafish FTP Test (Vari Export results Pasition Nearest Gene hr1= 85,5 dcunlo2 hr1= 799, gasti hr1= 217, hr1= 217, ags;154081 hr1= 724, zgs;56685	Type SNP SNP SNP SNP SNP SNP SNP SNP	Zyg. Gene Relation Het CE Hat CE Has CE Has CE	Variant g → a/g g → a/g g → a/g g → c/c g → a/h	rgdh: *Reads 5/4 2/2 3/18 4/4 5/5	Scare 5.52 18.65 27.60 12.26 14.96	Coding Changes Тумовузоиз [dcun142/EXSD4FT00000669371] тумовузоиз [dcun142/EXSD4FT0000069371] тумовузоиз [dcun142/EXSD4FT0000069371] тумовузоиз [dcun142/EXSD4FT00000693827] tyмовузоиз [dcun142/EXSD4FT00000693827] tymosysous [dcun132/EXSD4FT00000693827] tymosysous [dcun13265685/EXSD4FT00000693827]	Illeaf Score 25.00 44.00 25.00 25.00 25.00	cov. 9 4 21 4 6	Cone				

Cloud Computing Service



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
- 减少使用者終端的處理負擔
- 降低了使用者對於■「專業知識的依賴



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



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.

webservices Amazon Elastic Compute Cloud (Amazon EC2)

No More Gap





Features

Sequencing centers

Complete next-gen sequence data management

- On-demand compute and storage infrastructure
- Direct machine upload and LIMS integration
- Begin uploading first run in 10 minutes
- Data storage and distribution to users
- Run quality control
- Bioinformatics support for sequence analysis
- Illumina and SOLiD instrument support



Researchers

Next-gen sequence analysis and visualization

- On-demand compute and storage infrastructure
- Access everything through your web browser
- Long-term data storage and instant sharing
- Interactive genomic visualization
- RNA-Seq, ChIP-Seq, tag-based analysis
- Sequence variation detection
- Illumina, SOLiD, and Complete Genomics support

Learn more »



- | 不需軟體,上網即能分析與觀看結果
- 用電腦規格限制,雲端運算平行處理大量資料
- I 使core facility容易發佈數據,減低電腦成本、維護成本與分析負擔。
- I 使生物研究者能方便快速的自行整理NGS數據。
- 1 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



RNA-seq Analysis Pipeline Comparison



To Map or Not to Map...

Raw Reads (Before mapping) Raw Sequence Data

- •Fastq
- •Csfasta + QV

Mapped Reads (After mapping) Mapped (localized) SequenceSAM

•BAM

Fastq

@SEQ_ID

2 GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

!"*(((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>CCCCCC65

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

Line 2 is the raw sequence letters.

Illumina sequence identifiers

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

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



@ERR000451.1 VAB_S0103_20080915_542_14_17_70_F3
33023230203102103223330020300233001
%245719<.6353&:%0#\$1%&%2(--27*%&%,</pre>

csfastq

@ERR000451.2 VAB_S0103_20080915_542_14_17_171_F3
2332033212000200120221000000020001
#&#\$##&#&%%\$#&#%##'#&\$#%\$*&-))##%')

Sequence Alignment/Map (SAM)

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

<page-header><text><section-header><section-header><section-header>

- 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

Back

Upload reads file



DN∧nexus∭

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	<u>1474 Lib A</u>	1474 Lib B	<u>1474 Lib C</u>
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%
manned renetitively	9	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	925,371	936,460	29,691,554	23,943,881
confidently	5.12%	5.18%	54.95%	44.31%
Mapped	795,102	805,137	4,492,581	4,711,712
repetitively	4.40%	4.45%	8.31%	8.72%
Reads not	16,366,784	16,345,660	19,848,064	25,376,606
mapped	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
ribosomal RNA	15,767,755	15,392,529	307,188	287,866
	87.18%	85.10%	0.57%	0.53%
printer	1,000	1,000	7,112 0.019	27,010
control (phiX_174)	0.02%	0.01%	0.01%	0.03%
control (phix-1/4)	0 00%	0 00%	0 00%	0 00%
nolv-A	600	1 550	342	885
poiling	0 99 9 99%	a a1%	942 A AA%	885 0.00%
poly-C	204	975	0.00%	73
poly c	2 94 0 00%	975	92	9 99%
poly-G	196	224	0.00%	0.00%
poly G	0.00%	0.00%	0.00%	9 99%
	FIFL/A			

Quality control

Viewing statistics for 1 sample

Read statistics





Lab Users

Analyses

Q SEARCH CUSTOMIZE

CONDUCT NEW ANALYSIS

Listed below are all analyses to which you have access. Click on the analysis name to see analysis parameters and results. Select one ore more analyses to perform actions.

Filtering on keyword demo x

Analyses (10 of 23)

EXPORT RESULTS SHARE

DELETE

Select: All None

BROWSE RESULTS

÷	NAME	STATUS 🕈	OWNER 🕈		GENOME 🗢	TYPE 🗢	PERMISSIONS 🕈
	DEMO share *: Huge sample RNA-Seq	~	asimenos	Mar 17	H. sapiens (hg18)	RNA-Seq / Transcriptome-based quantification	View
	Demo Analysis: 3SEQ Regions RefSeq	1	dnanexus	Apr 16	H. sapiens (hg18)	3SEQ / Expressed regions in genome	Share
	Demo Analysis: ChIP Peaks RefSeq	~	dnanexus	Apr 16	H. sapiens (hg18)	ChIP-Seq / Peaks or regions	Share
	Demo Analysis: ChIP Regions RefSeq	~	dnanexus	Apr 16	M. musculus (mm9)	ChIP-Seq / Peaks or regions	Share
	Demo Analysis: RNA-Seq Brain RefSeq Human	~	dnanexus	Apr 16	H. sapiens (hg18)	RNA-Seq / Transcriptome-based quantification	Share
	Demo Analysis: RNA-Seq Brain RefSeq Mouse	4	dnanexus	Apr 16	M. musculus (mm9)	RNA-Seq / Transcriptome-based quantification	Share
	Demo Analysis: RNA-Seq Liver RefSeq Human	~	dnanexus	Apr 16	H. sapiens (hg18)	RNA-Seq / Transcriptome-based quantification	Share
	Demo Analysis: RNA-Seq Liver RefSeq Mouse	4	dnanexus	Apr 16	M. musculus (mm9)	RNA-Seq / Transcriptome-based quantification	Share
	Demo Analysis: RestrictionEnzyme; HCT116; HpaII	~	dnanexus	Apr 16	H. sapiens (hg18)	Restriction enzyme quantification	Share
	Demo Analysis: RestrictionEnzyme; HCT116; MspI	~	dnanexus	Apr 16	H. sapiens (hg18)	Restriction enzyme quantification	Share

Conduct a new analysis by selecting the experiment type and genome to see relevant samples.

Conduct new analysis

Analysis dei	ails									
Name:	My first analysis									
Experiment typ	ne / analysis: RNA-Seq / Trans	scr <mark>ipto</mark> r	me-based qua	ntification 💌	White paper					
Genome:	H. sapiens (hg18	3)	•							
Select samp	les to include in the analysis			Filter						
INCLUDE 🗢	NAME Demo Sample: Human liver total RNA	\$	CREATED = Mar 16	dnanexus	hg18 RNA-Seq	RUN NAME	•	RUN LANE 🔻	Share	+

RNA-seq Analysis

· · · · ·							
	chr11:74,793,0	87-74,793,582	Showing 496 bp	÷	$\rightarrow \ominus$	(9 10G 100M 1M
3,100	74,793,150	74,793,200	74,793,250 74	1,793,300 74	,793,350		74,793,400
	4	20	Plea	use zoom in to see data	for this track		1
$\rightarrow\rightarrow\rightarrow$	· · · · · · · · · · · · · · · · · · ·	$\rightarrow\rightarrow\rightarrow\rightarrow$	>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>	>>>>>RF	S3	>	>>>>
	SNO	RD15B	12222		official in the		
	-	-	-	-			
est splicing	alternative splici	ng data		0	-		
Export result	5		Read coun	t threshold			
Туре	Gene	Strand	Location	Read count 1 🔺			
► Junction	RPL23	-	chr17: 34,262,883-34,2	551.0			
▶ Junction	RPS14	÷	chr5: 149,807,490-149	556.4			
	0002	+	chr11: 74,788,268-74,	562.7			
▶ Junction	NP-55			CONTRACTOR OF A			
JunctionJunction	RPS29		chr14: 49,122,516-49,	679.7		¹ .	
 Junction Junction Junction 	RPS29 DYNLRB1	- +	chr14: 49,122,516-49, chr20: 32,577,808-32,	679.7 703.5			



ChIP-seq Analysis

	🤊 🛄 🇘	chr5: 137,828,424	4-137,829,535	Showing 1,111	bp	<>	00	10G 100M	1M 10K 1		0	Bookmark	Tracks	Option s	? Help
00	137,828,500	137,828,600	137,828,700	137,828,800	137,828,900	137,829,000	137	,829,100	137,82	9,200	137,82	9,3 0 0	137,829	9,400	137,829, I
					Please zoom in	to see data for thi	strack								
								EGR1	$\rangle\rangle\rangle$	$\langle \cdot \rangle$	>>	$\rangle \rangle$	$\rangle\rangle\rangle$	>>	>>>
								,,		N 3	10 S				
	BACK S	<u>}}</u> }	<u>\}</u>	$\langle / / / \rangle$	4 <u>/ </u>	` <i>\`\</i> `\							$\langle \gamma \rangle$	Ά,	
	17 1.]						Ϋ́								
N								M.							
							lemo A	nalysis: Ch	IP Peaks	RefSeq	peak da	ita - shov	wing 10,0	00 of 10,	000
							Export	t results	Search				log(qValue) Threshold	
							Rank	Position	Expt	Bkgd	Enri▲	-log(q\	Neares	Neares	Interg
							5	chr5: 137,	17,096	1.00	17,096	26,053			Yes
35.	000						4	chr7: 5,5	19,241	1.00	19,241	32,999	ACTB	ACTB	Yes
30,	000						3	chr5: 137,	22,892	1.00	22,892	46,710	EGR1	EGR1	Yes
25.	000		T				2	chr5: 137	26,906	1.00	26,906	64,530	EGR1	EGR1	Yes
20.	000						1	chr5: 60,1	40,401	1.00	40,401	145,49	ZSWIM6	ZSWIM6	Yes
15,	000														
10,	000						2.1								
5,0	00														



Mutation Analysis

49	92			Sho	wi	ng 5	о ы	p																Trac							
				3 13	21,1	154,4	460								21,1	154,4	470)							21,1	54,	480	l.			
6	G	G	С	С	A	A	С	С	С	А	G	Т	G	Т	Т	G	G	А	A	G	т	А	С	Т	С	Т	G	G	G	А	G
K	<	<	<	<	~	~	<	<	<	<	<	<	<	<	T	F	Y1	4	<	<	<	<	<	<	<	<	<	<	<	<	<
	$\langle \langle$	<	<	<	<	$\langle \langle$	<	<	\leq	\leq	<	<	K	\langle	<	<	<	C	Dź	24	Ŕ	<	<	\langle	<	<	Ś	Ś	<	Ż	Ś
- *	g	g	с	с	а	a	с	С	C	a	g	a	g	t	t	g	g	a	a	g	t	a	c	t	c	t	g	g	g	a	g
2 20	g	g	C	C	a	а	C	C	C	а	g	a	g	t	t	g	g	a	а	g	t	а	C	t	C	t	g	g	g	a	g
ho	g	g	С	С	а	а	С	С	С	а	g	а	g	t	t	g	g	а	а	g	t	а	С	t	С	t	g	g	g	а	g
P.	g	g	С	C	а	а	C	С	C	а	g	а	g	t	t	g	g	а	а	g	t	а	C	t	C	t	g	g	g	а	g
P	g	g	C	C	а	а	C	C	C	а	g	a	g	t	t	g	g	а	а	g	t	а	C	t	C	t	g	g	g	а	g
20	g	g	g	С	а	а	C																								
P	g	g	C	С	а	а	С	C	С	а	g	a	g	τ	τ	g	g	а	а	g	τ	а	C	τ	C	τ	g	g	g	а	g
11	iant	loc	i 01	/erl	арр	ing	COO	ling	exc	ons)	- 5	how	ing	27,9	20	of 2	7,9:	20												-	-
										Sea	arch																				
4 F	>			тур	be	Zy	/g.	1	Vari	iant		Sco	I IR	tef S	Sc.	Cov	<i>ı</i> .	#Re	ad	Cc	G	ene	13	Loc	atio	C	odin	g cl	na		
				SN	P	H	om		t ->	a/a		35.0	1	35.	00		46	31/	31			CD	24		CDS	т	-> 5	(CI	02+	•	
				SN	P	H	Dim.	- 4	c ->	t/t	-	28.5	5	35.	00		45	45/	45		SE	C61/	4.2		CDS	sy	nony	mou	s I	D	
				SN	P	Н	et	1	a ->	g/a	a	35.0	1	35.	00	1	51	21/	30			OR 58	32		CDS	v	-> /	(0)	RSE		
13	3/13		Autat	ion.p	sd [2	210.6	KB 6	525x4	84x2	4h ns	d 12	29%	Bicut	oic Int	erno	laton	Loz	aded in	0.0	5									-		1

NΛr	nexus	Home	Genome Brow	ser Sam	nples A	nalyses	Machin	es Lab Users		asundqui Sign o
Ana	alyses								Q SEARCH	
Listed	below are all analyses ses to perform actions.	to which you	u have access	. Click on t	he analysi	s name to	see ana	ysis parameters and r	esults. Select	one ore mor
Filteri	ng on keyword demo	ĸ								NEW ANALYSIS
Ana	alyses (10 of 23)									
BRO	WISE RESULTS EXPORT RESULTS	SI ARE DELET	E							
Sele	ct: All None									
÷	NAME		▲ STATUS ♥	OWNER 🕈	CREATED \$	GENOME	¢	ТҮРЕ	¢	PERMISSIONS
	DEMO share *: Huge sample	e RNA-Seq	4	asimenos	Mar 17	H. sapien:	s (hg18)	RNA-Seq / Transcriptome- quantification	-based	View
	Demo Analysis: 3SEQ Regio	ns RefSeq	1	dnanexus	Apr 16	H. sapiens	s (hg18)	3SEQ / Expressed regions in	n genome	Share
1	Demo Analysis: ChIP Peaks	RefSeq	1	dnanexus	Apr 16	H. sapiens	s (hg18)	ChIP-Seq / Peaks or region	15	Share
	Demo Analysis: ChIP Region	ns RefSeq	1	dnanexus	Apr 16	M. muscu (mm9)	ılus	ChIP-Seq / Peaks or region	15	Share
7	Demo Analysis: RNA-Seq Br	ain RefSeq Hum	an 🖌	dnanexus	Apr 16	H. sapiens	s (hg18)	RNA-Seq / Transcriptome- quantification	based	Share
	Demo Analysis: RNA-Seq Br	ain RefSeq Mou	se 🖌	dnanexus	Apr 16	M. muscu (mm9)	ilus	RNA-Seq / Transcriptome- quantification	-based	Share
	Demo Analysis: RNA-Seq Li	ver RefSeq Humi	an 🖌	dnanexus	Apr 16	H. sapien:	s (hg18)	RNA-Seq / Transcriptome- quantification	-based	Share
	Demo Analysis: RNA-Seq Li	ver RefSeq Mous	e 🗸	dnanexus	Apr 16	M. muscu (mm9)	ılus	RNA-Seq / Transcriptome- quantification	based	Share
	<u>Demo Analysis: RestrictionE</u> <u>HpaII</u>	inzyme; HCT116;	~	dnanexus	Apr 16	H. sapien	s (hg18)	Restriction enzyme quantif	fication	Share
	Demo Analysis: RestrictionE <u>MspI</u>	nzyme; HCT116;	1	dnanexus	Apr 16	H. sapien	s (hg18)	Restriction enzyme quantit	fication	Share
Sele	ct: <u>All None</u>	SHARE DELET	E							

	P - 19 - 19 -	•		exported_ar	alyses - Microsoft E	ixcel						X
9	Home In:	sert Pa	ge Layout Formulas Data	Review View	v Add-Ins						. 💿 -	. 🔹 >
Paste T	Calibri B Z rd 9	• 1 <u>∎</u> • [⊞ Font	$1 \cdot \begin{bmatrix} \mathbf{A}^* & \mathbf{A}^* \end{bmatrix} \equiv \equiv \equiv \bigotimes $	トー 日 Gene E 律 通子 S ・ nt G	ral ← % • (Conditiona Formatting	I Format * as Table * Styles	Cell Styles *	G a Insert → Cells	Σ · A · Z · Z · Sor · Filto Ed	t & Find & er * Select *	
A768 - f AHCYL1												
	A	В	С	D	E	F	G	Н	1	L	К	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			1			
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					1	
774	AHDC1			chr1:27803238	104.3	949.35						
775	AHI1	4	chr6:135646804-135860595	chr6:135860598	158.1	2199.08						
776		12 analyses	chr11.62039949-62070907	chr11.62066998	60.7	216 54			III			
Ready	- exported				Average:	471.17 Cou	unt: 10 S	um: 2355.8	5 🔳 💷	100% (-)		(+







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.

		% "best mapping" incorrect				% true source not found				
method	reads/s	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	<mark>4.56</mark>	8.66	<mark>23.19</mark>	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
1. N	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 = Σ 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

Туре	Description	Some examples (Reference -> Genotype) and their meanings
SNP	A single nucleotide is	A -> A/G (heterozygous SNP)
	mutated	A -> G/G (homozygous SNP)
		A -> G/T (heterozygous SNP, new alleles)
MNP	A series of consecutive	AA -> AA/CT (heterozygous MNP)
	nucleotides are mutated	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	A -> /C (deletion and mutation)
	events	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 <u>BED file format</u>. 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.