RNA-Seq実験ノート:
リード長とリード数のデザインとウェット実験の注意点

東大・新領域

鈴木 穣

ysuzuki@hgc.jp

TOPICS:

RNAを鋳型に用いた次世代シークエンス解析

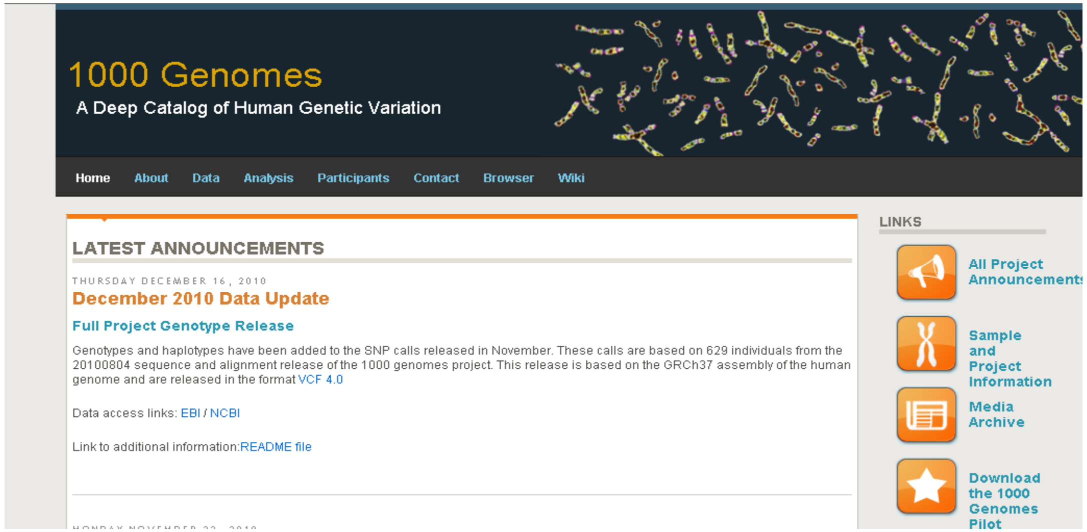
RNA Seqの一般論：よくある問題点

アプリケーションあれこれ (mRNA, miRNA, RIP Seq)
~実験のデザイン->何でもやったらいいのでは・・・

多検体処理：鋳型調整の自動化を背景に

Sequence production rate acceralated

Whole Genome Sequnecing in “1000 Genome Project”



1000 Genomes

A Deep Catalog of Human Genetic Variation

LATEST ANNOUNCEMENTS

THURSDAY DECEMBER 16, 2010

December 2010 Data Update

Full Project Genotype Release

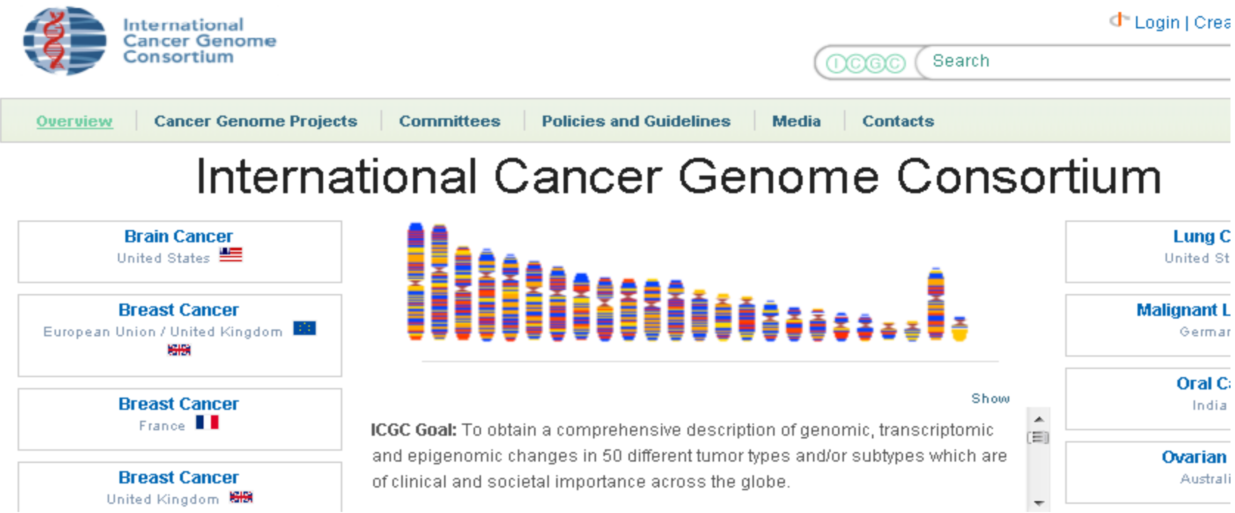
Genotypes and haplotypes have been added to the SNP calls released in November. These calls are based on 629 individuals from the 20100804 sequence and alignment release of the 1000 genomes project. This release is based on the GRCh37 assembly of the human genome and are released in the format VCF 4.0

Data access links: EBI / NCBI

Link to additional information:README file

LINKS

“Exome Sequencing in Cancer genome project



International Cancer Genome Consortium

Brain Cancer
United States

Breast Cancer
European Union / United Kingdom

Breast Cancer
France

Breast Cancer
United Kingdom

Lung C
United St

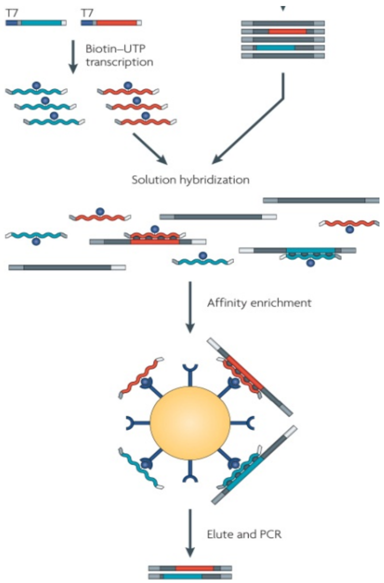
Malignant L
Germar

Oral C
India

Ovarian
Australi

ICGC Goal: To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

“Exon capture”



Improved sequencing capacity of “Next Generation” reduced the time and cost for the genome/cDNA analysis

Wash-U Cancer Re-sequencing design

Whole Genomic DNA (Cancer: 50X)

Whole Genomic DNA (Normal; 50X)

Exome (Cancer)

Exome (Normal)

RNA Seq (Cancer)

1 run

変異遺伝子発現の有無

転座遺伝子の有無

（遺伝子発現変化）

“Sample Prep automation” is necessary even in a small setup

Sequence facility in Univ. Tokyo, Kashiwa

GAIIx 3+ Hiseq2000 1@The Univ. of Tokyo

Operation:

Technicians 4

“ゲノム支援”

「ゲノム支援」はゲノム科学の
すそ野拡大とピーク作りをめざします

最新情報

2010.09.01 平成22年度第1回公募は締め切りました。167件の申請がありましたが、支援課題の選定結果は9月末に申請者に連絡する予定です。

2010.08.09 情報解析支援活動の支援可能な内容を修正しました。

2010.08.06 説明同意文書（インフォームドコンセント）のモデル書式の中で同意文書部分が抜けておりましたが、追加しました。

2010.08.05 FAQを更新しました。

2010.08.03 支援課題の公募要領、及び、FAQを掲載しました。

2010.07.30 平成22年度第一回支援課題公募は、8月3日（火）から申請受付開始、8月31日（火）締めきりの予定です。支援申請等はすべてこのホームページ上で行う予定です

Providing NGS platform for researchers in various research fields

http://www.genome-sci.jp/

			胎児型腎臓幹細胞の成体腎での再活性化
			次世代シークエンサーを用いた生殖系列のエピゲノム修飾とトランスクリプトーム解析
		5	種内雑種を利用した対立遺伝子間の優劣に関わるDNAメチル化機構の解析
			メリステム制御の基盤を支える植物幹細胞の不等分裂の分子機構の解明
			トゲウオ科魚類における種分化の遺伝機構
藤堂 剛	大阪大学		メダカ逆遺伝学的手法を基盤とした個体・組織レベルでの損傷応答解析系の確立
太田 邦史	東京大学	8	長鎖非翻訳RNAを介したクロマチン／染色体機能の制御
武田 洋幸（森下BS）	東京大学		組織が創るマクロでロバストなコンパートメントの成立・維持のロジック
深田 吉孝	東京大学		脳時計ニューロンにおける光シグナリングと概日リズム制御の分子解析
多羽田 哲也	東京大学		ショウジョウバエの記憶形成回路の構造および機能発現の分子基盤
三谷 啓志	東京大学		個体内における電離放射線誘発突然変異成立過程の解明
平良 眞規	東京大学		転写制御ネットワークから見る原口形成と原腸胚オーガナイザーの進化のメカニズム
國枝 武和	東京大学		極限環境耐性動物クマムシが獲得した耐性メカニズムの解明
稲田 利文	名古屋大学理学研究科		新生ポリペプチド鎖依存の翻訳アレストにおけるRACK1の機能解明
高浜 洋介	徳島大学		胸腺における自己形成と自己認識
嶋田 透	東京大学		カイコとその近縁種における寄主植物選択機構の進化
田中 知明	千葉大学		p53転写因子複合体によるクロマチン機能調節とiPSリプログラム制御機構の解明
後藤 由季子	東京大学		胎生期大脳新皮質神経幹細胞による多様な細胞の産生機構の解析
坂山 英俊	神戸大学		陸上植物の2倍体多細胞体制の起源をシャジクモ藻類の遺伝子から探る
三室 仁美	東京大学		ヘリコバクターピロリの胃粘膜感染機構と炎症惹起メカニズムの研究
國府 力	大阪大学		初期発生におけるクロマチン制御のリアルタイム解析
田中 知明	千葉大学		転写因子p53による新たな代謝調節機能と代謝環境応答のエピジェネティクス制御
福澤 秀哉	京都大学		デジタル遺伝子発現解析による微細藻類のCO2濃縮・水素発生関連遺伝子の同定と

RNA Seqの分類

タグ数をカウントするもの　(36bp Single End Read)

参照ゲノム配列あり

- 発現量を計測するもの
  - (mRNA) RNA Seq
  - small RNA Seq
- タンパク質との相互作用を計測するもの　RIP Seq/CLIP Seq

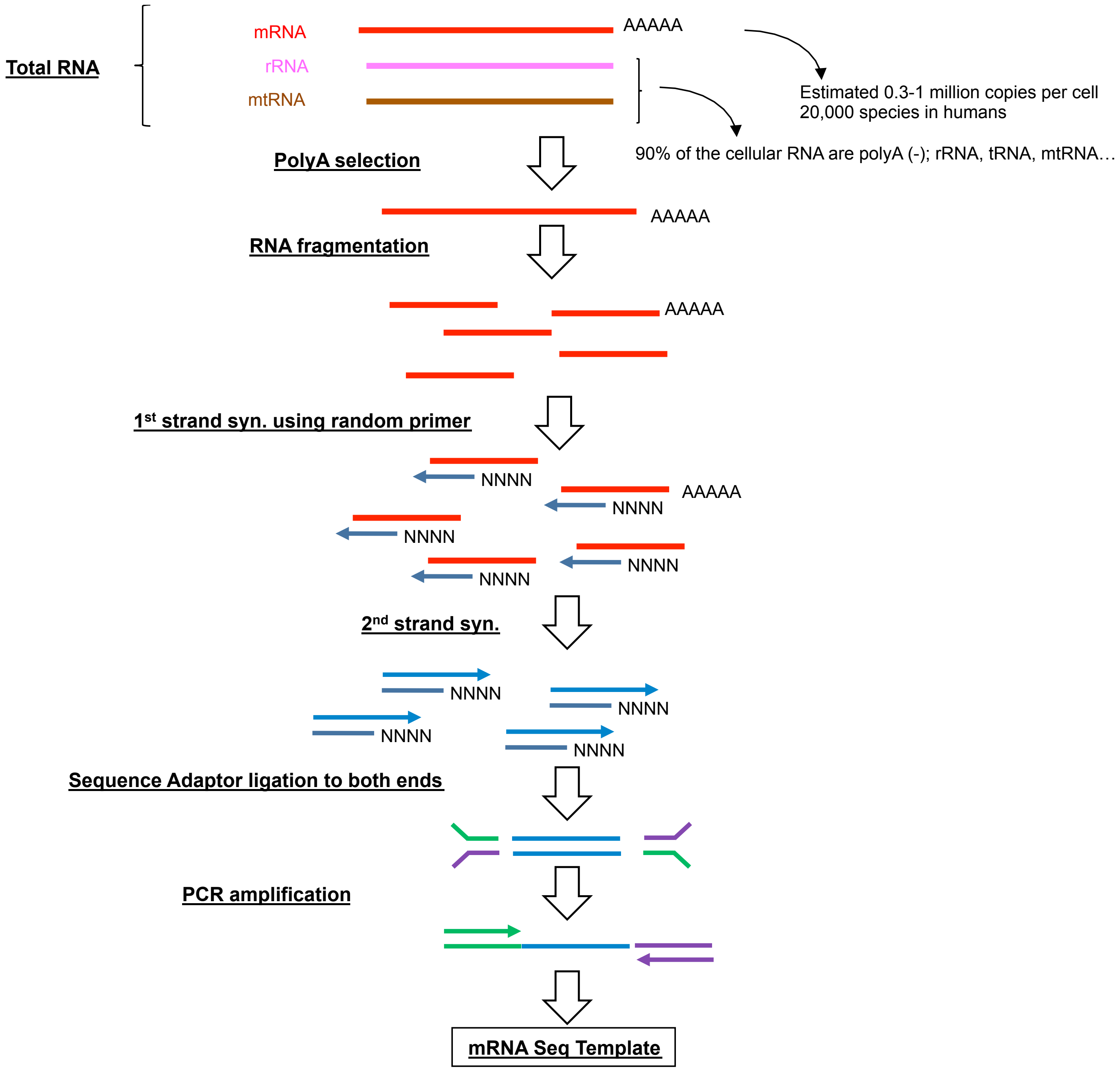
配列を決定するもの (>76bp Paired End Read)

参照ゲノム配列なし

- 遺伝子アノテーションするもの　de novo アセンブリ
- 選択的スプライシングを解析するもの

mRNA Seq

Template Prep. for RNA Seq



Total RNA

Estimated 0.3-1 million copies per cell
20,000 species in humans

90% of the cellular RNA are polyA (-); rRNA, tRNA, mtRNA…

PolyA selection

RNA fragmentation

1st strand syn. using random primer

2nd strand syn.

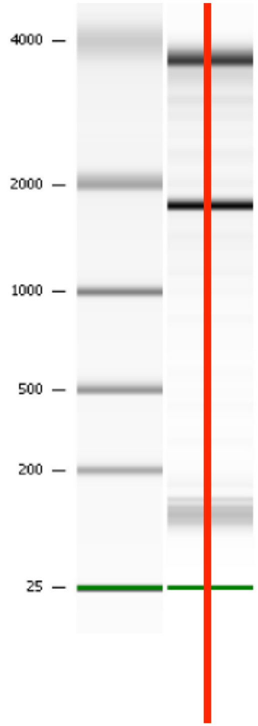
Sequence Adaptor ligation to both ends

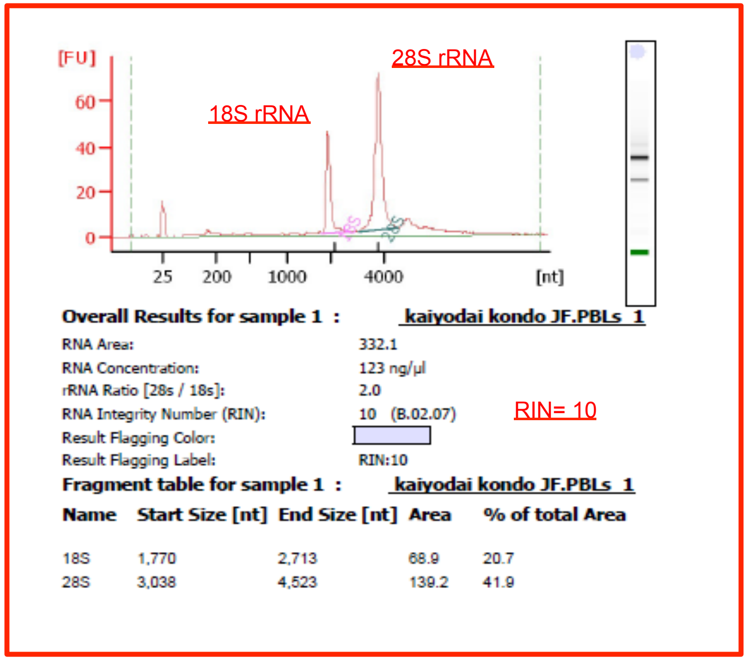
PCR amplification

mRNA Seq Template

BioAnalyzer is essential for sample preparation

BioAnalyzer (Agilent):
Electrophoresis on microchip





Overall Results for sample 1 : **kaiyodai kondo JF.PBLs 1**

RNA Area: 332.1
RNA Concentration: 123 ng/µl
rRNA Ratio [28s / 18s]: 2.0
RNA Integrity Number (RIN): 10 (B.02.07)
Result Flagging Color:
Result Flagging Label: RIN:10

RIN= 10

Fragment table for sample 1 : **kaiyodai kondo JF.PBLs 1**

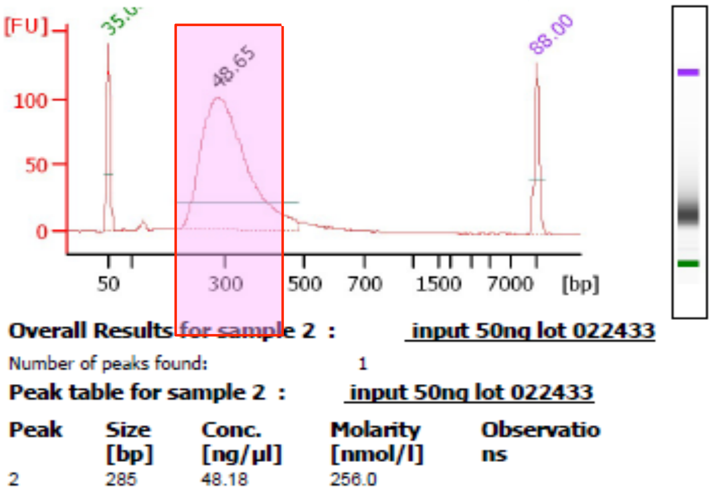
Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,770	2,713	68.9	20.7
28S	3,038	4,523	139.2	41.9

Dissection

※別のメーカーの宣伝ですいません・・・

Advantages in using BioAnalyzer (I)

effective material (250-450 bp)



Overall Results for sample 2 : **input 50ng lot 022433**

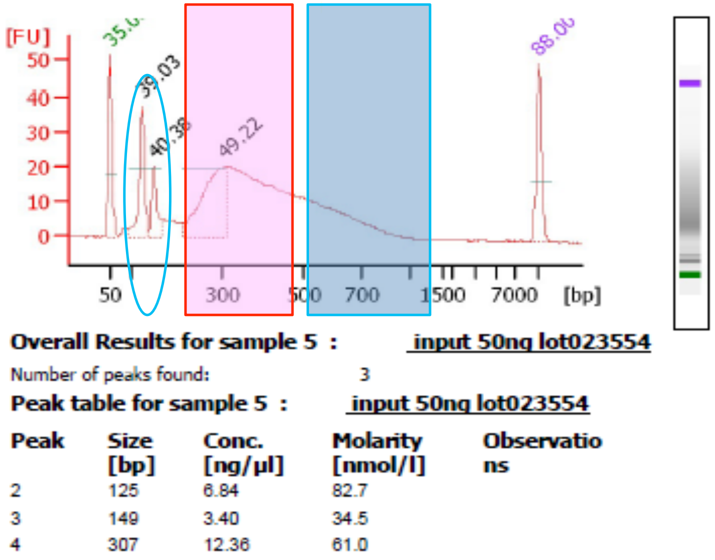
Number of peaks found: 1

Peak table for sample 2 : **input 50ng lot 022433**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	285	48.18	256.0	

effective material (250-450 bp) **non-effective material**

Primer dimer



Overall Results for sample 5 : **input 50ng lot023554**

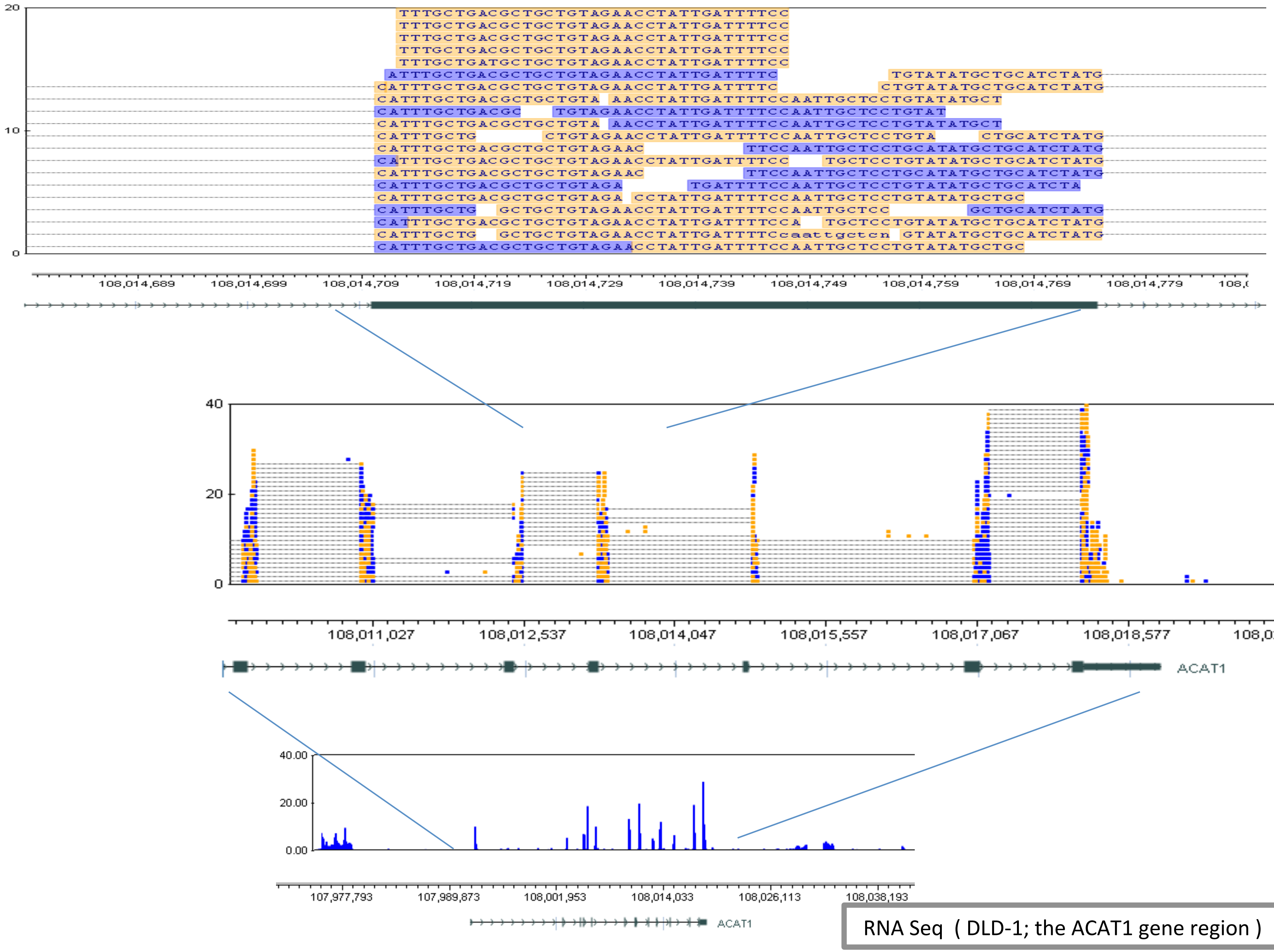
Number of peaks found: 3

Peak table for sample 5 : **input 50ng lot023554**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	125	6.84	82.7	
3	149	3.40	34.5	
4	307	12.36	61.0	

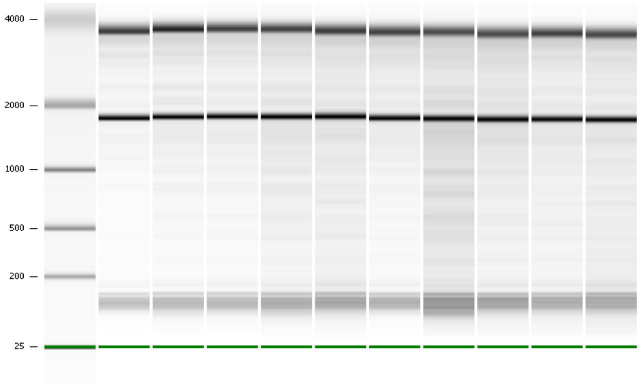
To measure effective template amount

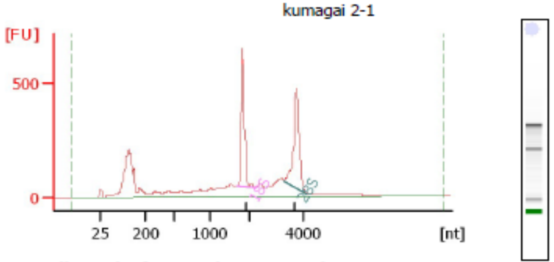
Examples of NGS data (RNA Seq on Genome Studio Viewer)



RNA Seq (DLD-1; the ACAT1 gene region)

Sample Prep for Time-Course RNA Seq Analysis

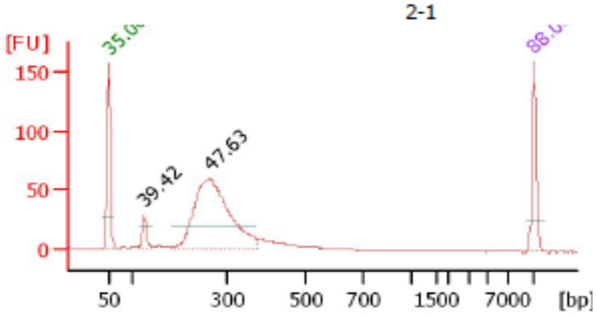


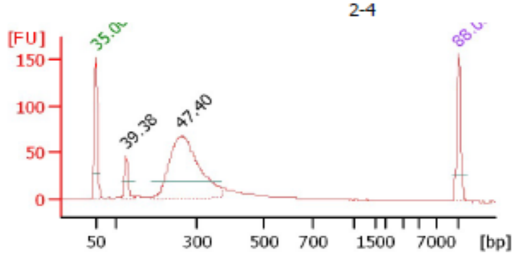


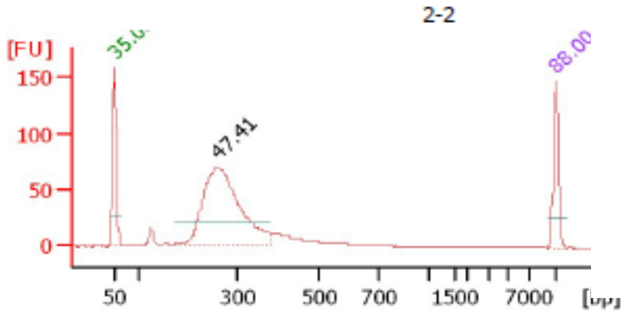
Overall Results for sample 6 :	**kumagai 2-1**
RNA Area:	4,593.5
RNA Concentration:	25,998 pg/µl
rRNA Ratio [28s / 18s]:	1.2
RNA Integrity Number (RIN):	7.8 (B.02.07)
Result Flagging Color:	🟪
Result Flagging Label:	RIN: 7.80

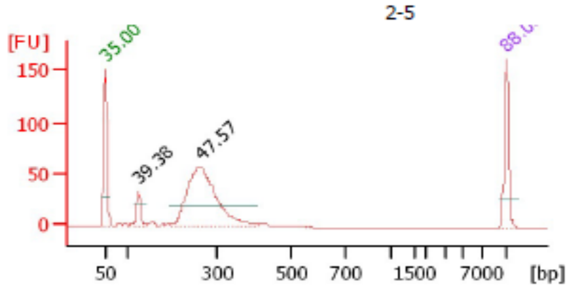
Fragment table for sample 6 : **kumagai 2-1**

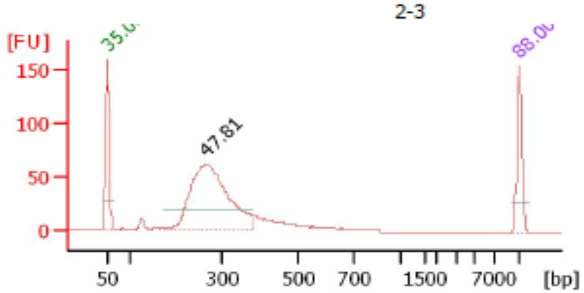
Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,647	2,209	642.2	14.0
28S	3,244	4,125	773.5	16.8







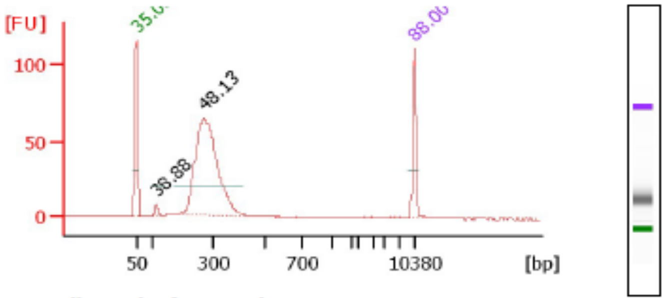




Starting material: total RNA >1microG (実際は>100ng -> 10ng)

Various types of Sample Preparation

Human Colon Cancer RNA



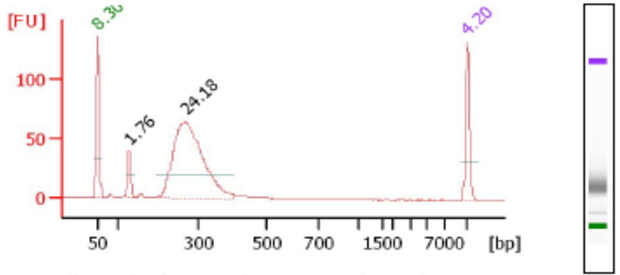
Overall Results for sample 1 : **mimori 33**

Number of peaks found: 2

Peak table for sample 1 : **mimori 33**

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	116	0.68	8.9	
3	274	26.69	147.7	

Flatfish RNA



Overall Results for sample 1 : **kaiyoukenn**

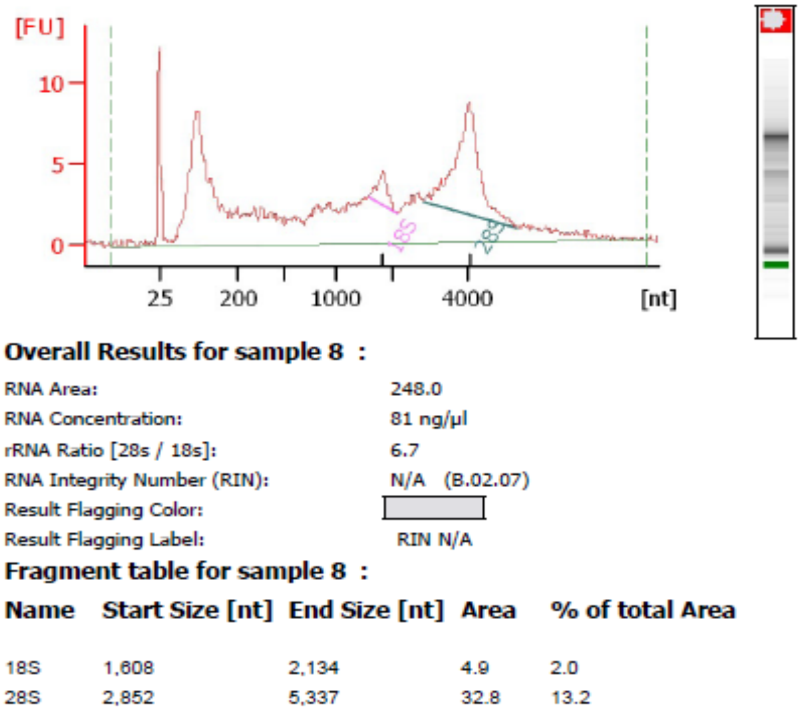
Number of peaks found: 2

Peak table for sample 1 : **kaiyoukenn**

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	50	8.30	251.5	Lower Marker
2		127	1.76	21.0	
3		266	24.18	137.6	
4	▶	10,380	4.20	0.6	Upper Marker

Occasionally, “irregular samples” should be also handled

Total RNA from operation material



Overall Results for sample 8 :

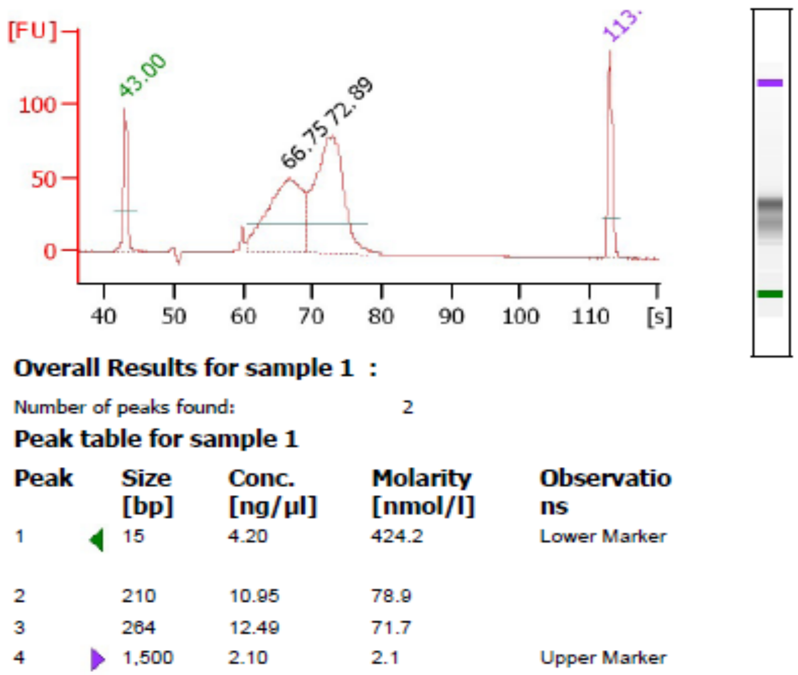
RNA Area: 248.0
RNA Concentration: 81 ng/µl
rRNA Ratio [28s / 18s]: 6.7
RNA Integrity Number (RIN): N/A (B.02.07)
Result Flagging Color: ▨
Result Flagging Label: RIN N/A

Fragment table for sample 8 :

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	1,608	2,134	4.9	2.0
28S	2,852	5,337	32.8	13.2

RIN N/A; but this is still RNA!

“irregular” template



Overall Results for sample 1 :

Number of peaks found: 2

Peak table for sample 1

Peak		Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	◀	15	4.20	424.2	Lower Marker
2		210	10.95	78.9	
3		264	12.49	71.7	
4	▶	1,500	2.10	2.1	Upper Marker

（おもに）ゲノムアノテーション /遺伝子発見

ある魚類のdenovo

● data process

Illumina Read 76PE
(Pass Filtered , remove the read including N)

↓

AbySS (version 1.2.6)

↓

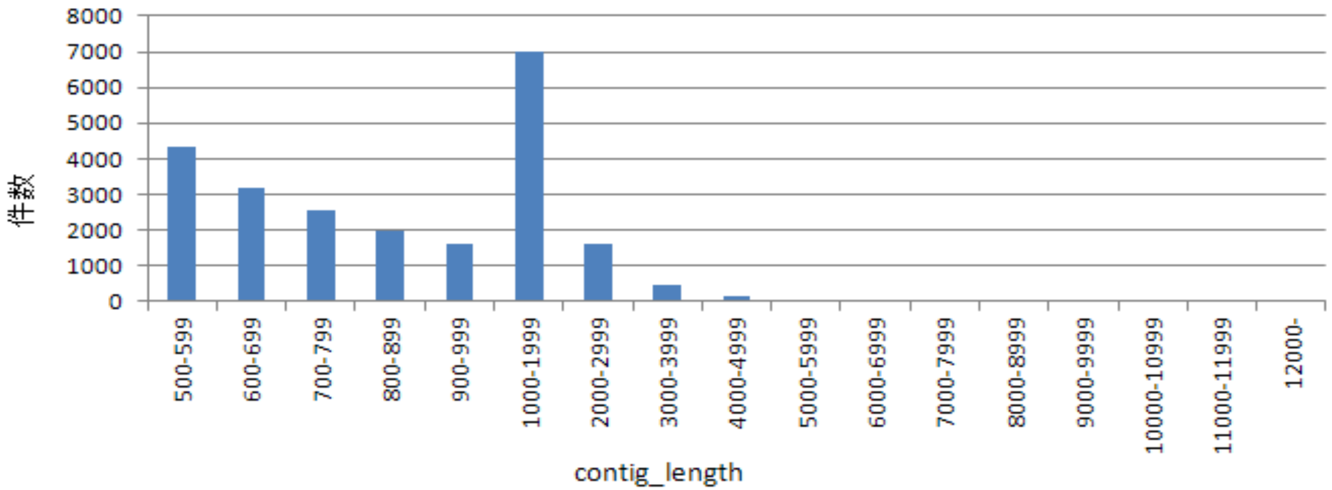
> 500bp contig 抽出

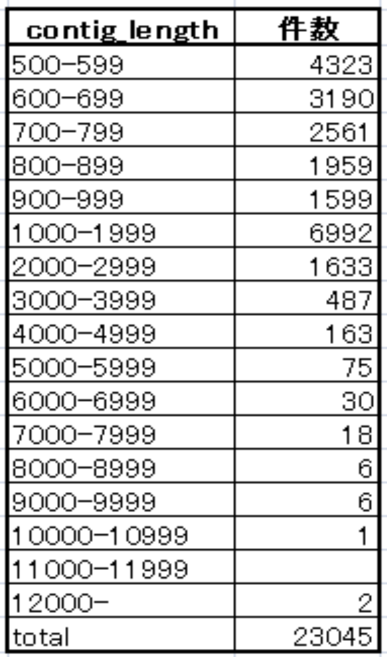
↓ ↘

tBlastX (Query:contig , DB: NT)   ELAND (Ref:contig)

●assemble result

Sample	# Reads (76bp)	# Assembled contigs 500bp< Average contig length	#Matched with tBLASTX < 1e-50 500bp<
JDPBLs-1	46,771,912	23,045 （Average 1,141bp)	11,549



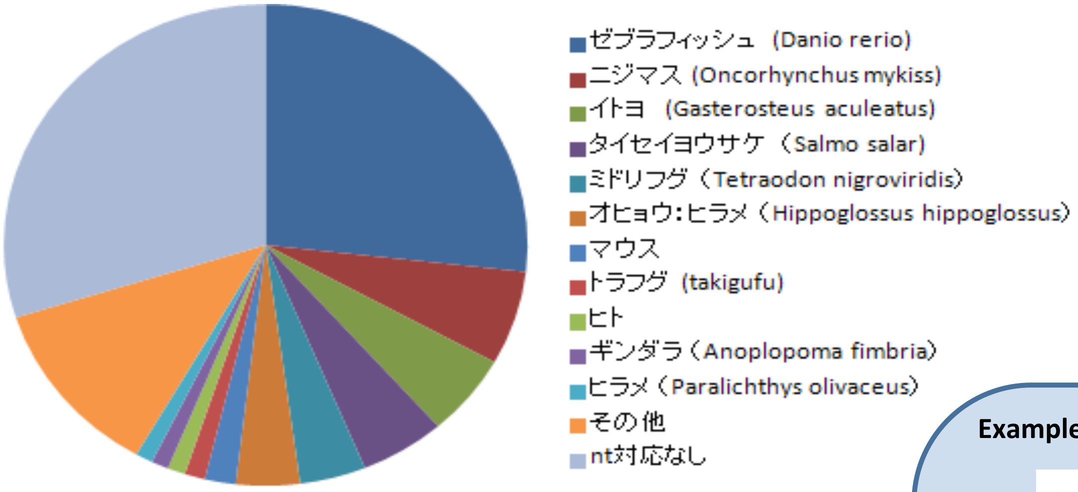


contig_length	件数
500−599	4323
600−699	3190
700−799	2561
800−899	1959
900−999	1599
1000−1999	6992
2000−2999	1633
3000−3999	487
4000−4999	163
5000−5999	75
6000−6999	30
7000−7999	18
8000−8999	6
9000−9999	6
10000−10999	1
11000−11999	
12000−	2
total	23045

近藤研との共同研究

ある魚類のdenovo

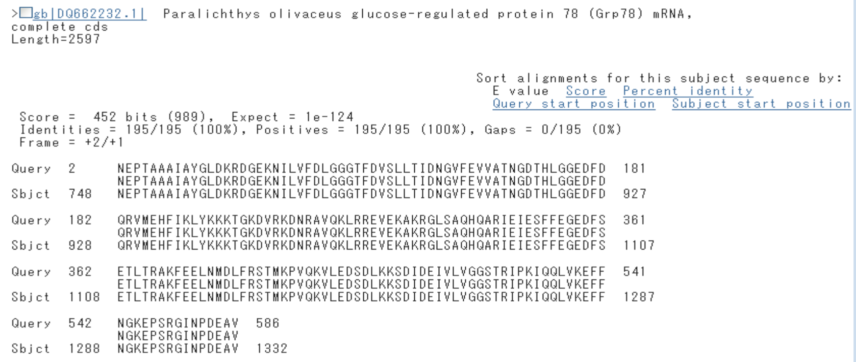
●tblastx assembled contig to NT



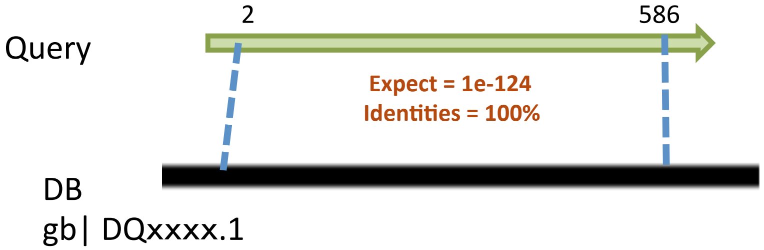
tblastx結果 内訳	
ゼブラフィッシュ（*Danio rerio*）	27％
ニジマス（*Oncorhynchus mykiss*）	6％
イトヨ （*Gasterosteus aculeatus*）	5％
タイセイヨウサケ（*Salmo salar*）	5％
ミドリフグ（*Tetraodon nigroviridis*）	4％
オヒョウ:ヒラメ（*Hippoglossus hippoglossus*）	4％
マウス	2％
トラフグ（*takigufu*）	1％
ヒト	1％
ギンダラ（*Anoplopoma fimbria*）	1％
ヒラメ（Paralichthys olivaceus）	1％
その他	12％
nt対応なし	30％

Example: xxx Assembled contig : Query length 588bp

```
>contig_102559 588 97855
CAATGAGCCAACTGCTGCTGCCATTGCTTATGGTCTGGACAAGAGAGATGGCGAGAAGAACATTCTTGTGTTCGATCT
GGGTGGCGGCACCTTCGATGTCTCCCTCTTGACCATCGACAATGGTGTGTTTGAAGTGGTGGCCACCAACGGTGACAC
TCACCTGGGAGGTGAGGACTTCGACCAGCGCGTCATGGAGCACTTCATCAAGCTGTACAAGAAGAAAACTGGCAAAG
ATGTGCGCAAAGACAACCGTGCTGTGCAGAAGCTGCGTCGTGAGGTTGAGAAGGCAAAGAGGGGGCTGTCCGCCCA
GCACCAGGCCCGCATTGAGATCGAGTCCTTCTTTGAGGGAGAAGACTTCTCTGAGACTCTGACCCGTGCCAAGTTTGA
AGAGCTGAACATGGACCTGTTCCGTTCCACCATGAAGCCTGTGCAGAAGGTGCTGGAAGATTCCGACCTGAAGAAAT
CTGACATCGATGAGATTGTCCTGGTTGGAGGCTCCACCCGTATCCCCAAAATTCAGCAGCTGGTGAAGGAGTTCTTCA
ATGGCAAGGAGCCATCTAGGGGCATCAACCCTGATGAGGCTGTGGC
```



```
>☐gb|DQ662232.1|  Paralichthys olivaceus glucose-regulated protein 78 (Grp78) mRNA,
complete cds
Length=2597


                                        Sort alignments for this subject sequence by:
                                          E value  Score  Percent identity
                                          Query start position  Subject start position
 Score =  452 bits (989),  Expect = 1e-124
 Identities = 195/195 (100%), Positives = 195/195 (100%), Gaps = 0/195 (0%)
 Frame = +2/+1

Query  2     NEPTAAAIAYGLDKRDGEKNILVFDLGGGTFDVSLLTIDNGVFEVVATNGDTHLGGEDFD  181
             NEPTAAAIAYGLDKRDGEKNILVFDLGGGTFDVSLLTIDNGVFEVVATNGDTHLGGEDFD
Sbjct  748   NEPTAAAIAYGLDKRDGEKNILVFDLGGGTFDVSLLTIDNGVFEVVATNGDTHLGGEDFD  927

Query  182   QRVMEHFIKLYKKKTGKDVRKDNRAVQKLRREVEKAKRGLSAQHQARIEIESFFEGEDFS  361
             QRVMEHFIKLYKKKTGKDVRKDNRAVQKLRREVEKAKRGLSAQHQARIEIESFFEGEDFS
Sbjct  928   QRVMEHFIKLYKKKTGKDVRKDNRAVQKLRREVEKAKRGLSAQHQARIEIESFFEGEDFS  1107

Query  362   ETLTRAKFEELNMDLFRSTMKPVQKVLEDSDLKKSDIDEIVLVGGSTRIPKIQQLVKEFF  541
             ETLTRAKFEELNMDLFRSTMKPVQKVLEDSDLKKSDIDEIVLVGGSTRIPKIQQLVKEFF
Sbjct  1108  ETLTRAKFEELNMDLFRSTMKPVQKVLEDSDLKKSDIDEIVLVGGSTRIPKIQQLVKEFF  1287

Query  542   NGKEPSRGINPDEAV  586
             NGKEPSRGINPDEAV
Sbjct  1288  NGKEPSRGINPDEAV  1332
```



contig領域 タグ集計		
tag	ppm	rpkm
2035	94	159.86

Chinese hamster ovary denovo

● data process

Illumina Read 76PE
(Pass Filtered , remove the read including N)

↓

AbySS (version 1.2.6)

↓

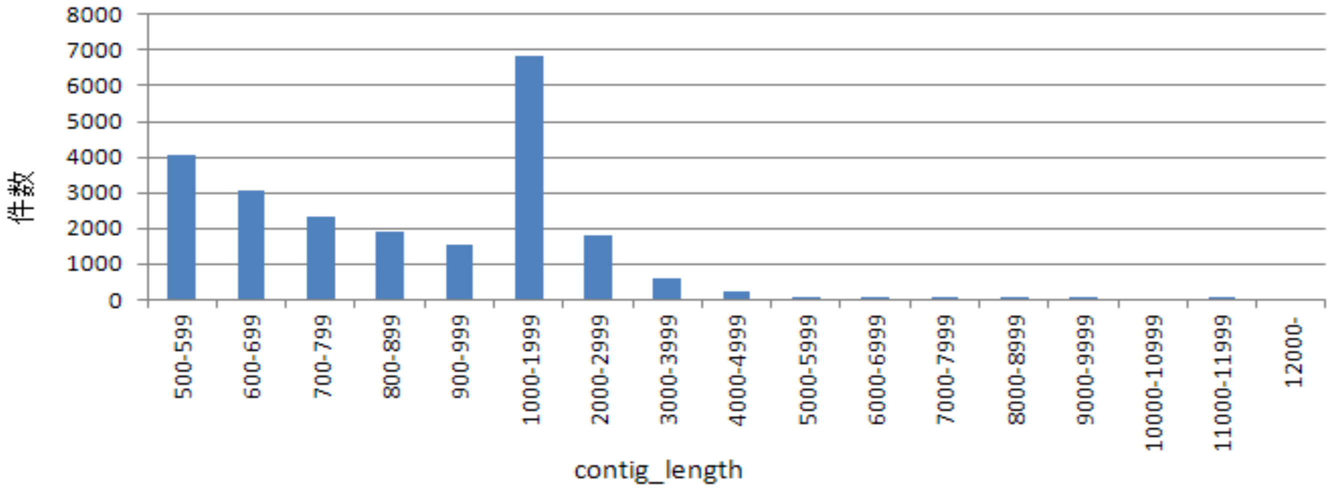
> 500bp contig 抽出

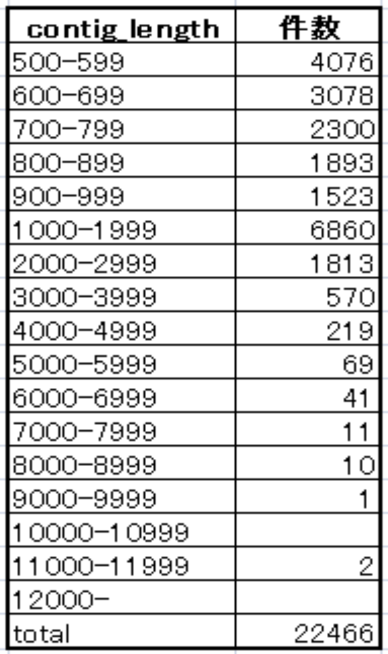
↓ ↘

tBlastX (Query:contig , DB: NT)   ELAND (Ref:contig)

●assemble result

Sample	# Reads (76bp)	# Assembled contigs 500bp< Average contig length	#Matched with tBLASTX < 1e-50 500bp<
CHO total RNA-1	42,552,668	22,466 （Average 1,184bp)	17,108
CHO total RNA-2	51,249,176	24,833 (Average 1,149bp)	18,070



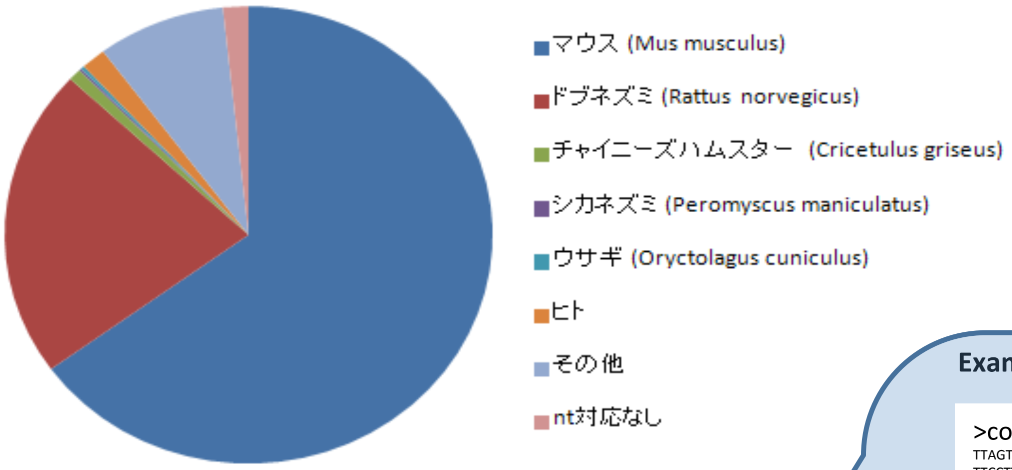


contig_length	件数
500−599	4076
600−699	3078
700−799	2300
800−899	1893
900−999	1523
1000−1999	6860
2000−2999	1813
3000−3999	570
4000−4999	219
5000−5999	69
6000−6999	41
7000−7999	11
8000−8999	10
9000−9999	1
10000−10999	
11000−11999	2
12000−	
total	22466

副島研との共同研究

Chinese hamster ovary　denovo

●tblastx assembled contig to NT

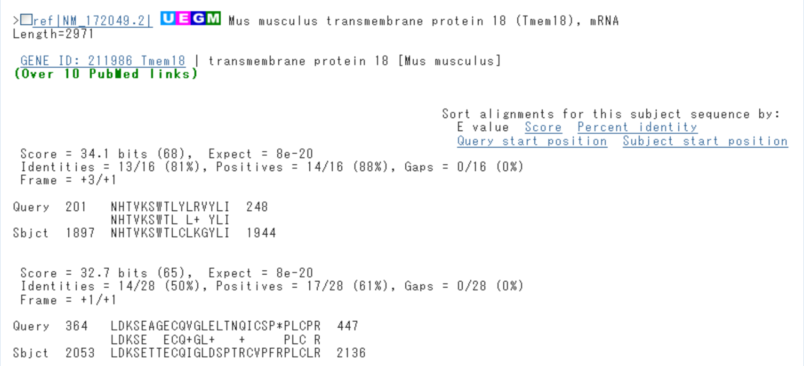


tblastx結果 内訳

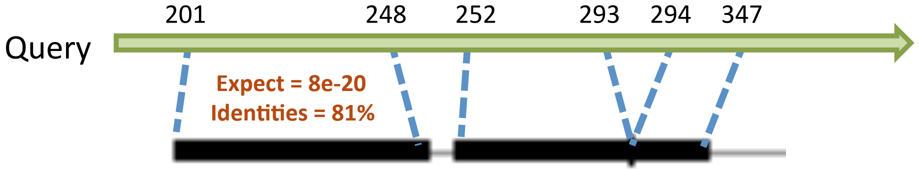
マウス（*Mus musculus*）	66%
ドブネズミ（*Rattus norvegicus*）	22%
チャイニーズハムスター（*Cricetulus griseus*）	1%
シカネズミ（*Peromyscus maniculatus*）	0%
ウサギ（*Oryctolagus cuniculus*）	0%
ヒト	2%
その他	9%
nt対応なし	2%

Example: CHO Assembled contig : Query length 652bp

```
>contig_243_652_3342
TTAGTAGAAAGCCTACCGACTGAATGGCCACAGTCTGATAATCCTAAAGTGTGAATCTGTGG
TTCCTTCATGAAGCCGCTGCAGCACTTAACCTGAATTGTGCTCTCAAAGTCTTGATTTGTTCCA
CTTTTACACTGAACAACAATAGCAAAAGTACCAGCAGAAATAACATAGTATAGTAAGTCAAA
TTTTAAATTTAAAACCACACTGTTAAGTCATGGACATTGTATCTAAGGGTATATCTAATTTGTG
TTTCTGAAACTACTGGTTCAGGGTGTTTAATCATTTTGTTTACGAAGGCACTCATGACAACATT
GATTAGTTTCACAGAAAATGACAATATATTGTTTAGTTATATTTAGTTAGATAAATCTGAAGCA
GGAGAGTGTCAGGTAGGTCTAGAACTCACCAATCAGATCTGTTCCCCTTGACCTTTGTGTCCA
AGGATATTGGTCCATATAGGAAAGAGAGTGGTTGAGCTGCTTGAGTTCTGGCATTGTCCAGA
GCAGCTTCAGGGGTGATGGAGAAACTAGAAAGAGAGGGGACTTGTAGATTCTGTGCATTTT
GAGACAACACTATCACTAATCATACATTGTTTAGACGGACTGTATTTGGAACCATCCAACTTA
ATCTAGTTTTTTTCTCTGATAA
```



```
>☐ref|NM_172049.2| UEGM Mus musculus transmembrane protein 18 (Tmem18), mRNA
Length=2971

 GENE ID: 211986 Tmem18 | transmembrane protein 18 [Mus musculus]
(Over 10 PubMed links)


                                        Sort alignments for this subject sequence by:
                                          E value  Score  Percent identity
                                          Query start position  Subject start position
 Score = 34.1 bits (68),  Expect = 8e-20
 Identities = 13/16 (81%), Positives = 14/16 (88%), Gaps = 0/16 (0%)
 Frame = +3/+1

Query  201   NHTVKSWTLYLRVYLI  248
             NHTVKSWTL L+ YLI
Sbjct  1897  NHTVKSWTLCLKGYLI  1944


 Score = 32.7 bits (65),  Expect = 8e-20
 Identities = 14/28 (50%), Positives = 17/28 (61%), Gaps = 0/28 (0%)
 Frame = +1/+1

Query  364   LDKSEAGECQVGLELTNQICSP*PLCPR  447
             LDKSE  ECQ+GL+   +     PLC R
Sbjct  2053  LDKSETTECQIGLDSPTRCVPFRPLCLR  2136
```



DB
ref| NM_172049.2

contig領域 タグ集計

tag	ppm	rpkm
71	4.05	6.21

（おもに）mRNA の解析
=>sub populationのmRNAの解析

Variation in Expression Levels between Patients

A Gene of Similar Expression Level

Expression level

Sequence is too dense for display.

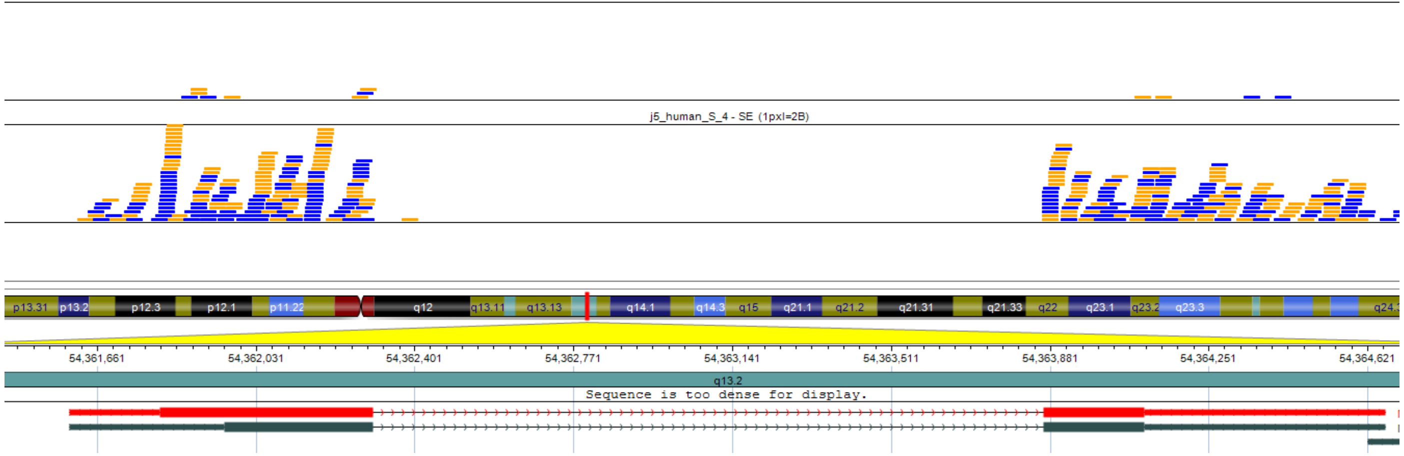
Patient I

Patient II

Gene model

A Gene of Different Expression Level

Expression level

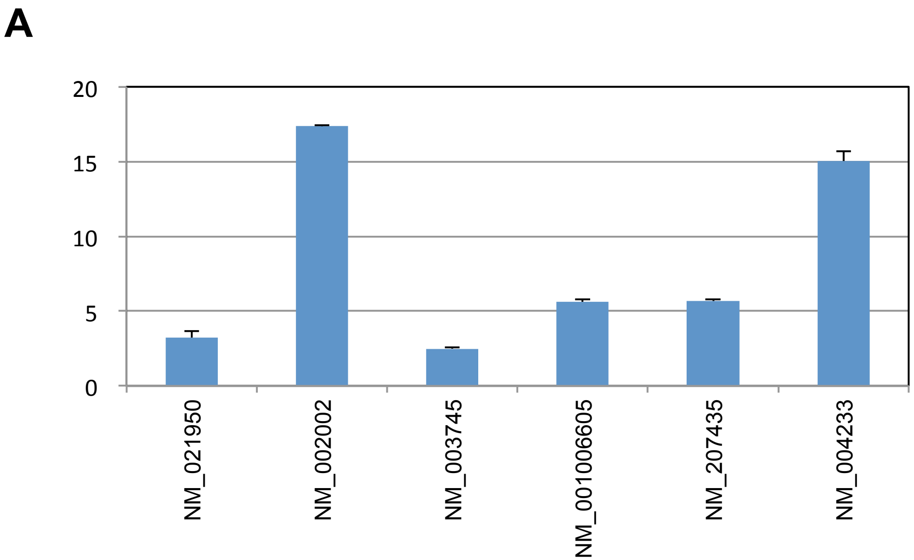


Sequence is too dense for display.

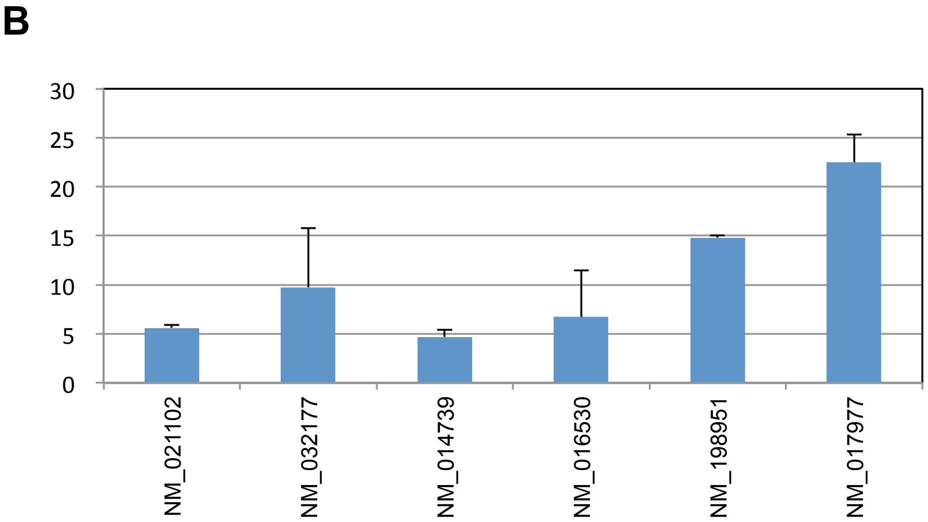
Patient I

Patient II

Gene model



TSSseq Fold	**4.9**	**7.9**	**9.1**	**6.1**	**11.9**	**13.7**



TSSseq Fold	**2.8**	**6.1**	**28.4**	**13.2**	**15.0**	**26.0**

Fold change in gene expression (stimulation +/-)

	AK025835	AK092486	AK097884
Tag fold (tag count)	5.5 (183/32)	8.7 (175/20)	6.3 (11/2)

Real Time RT-PCR fold: Primers set at different positions

AK025835	AK092486	AK097884
1.42	6.26	4.65
2.23	*4.99	1.32
2.05	11.54	2.44
1.59	1.03	*1.29
1.40	4.65	*3.02
*1.91		1.53
1.88		5.12
		*22.92

*: Ct>35 cycle

N=3 for each

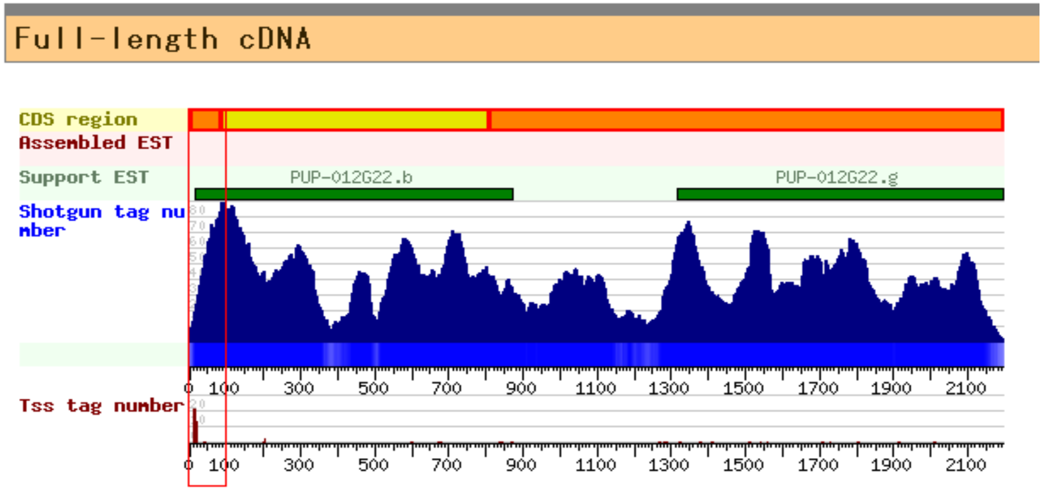
Stadard dev in each primer is mostly <10%

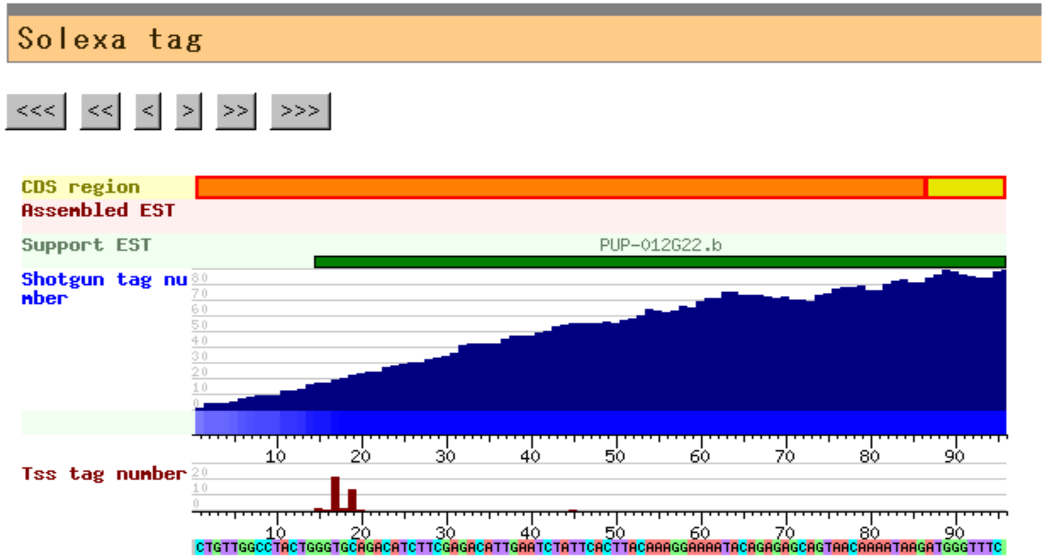
Even within RT-PCR, which RT-PCR results should be used?

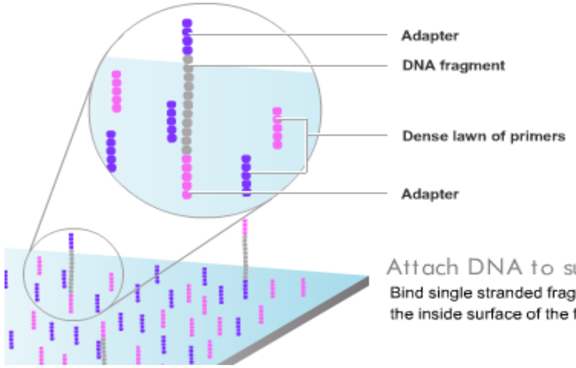
Biased tag distribution

Fragmentation efficiency?
Preference in sequence?

Bias in Expression Profile???

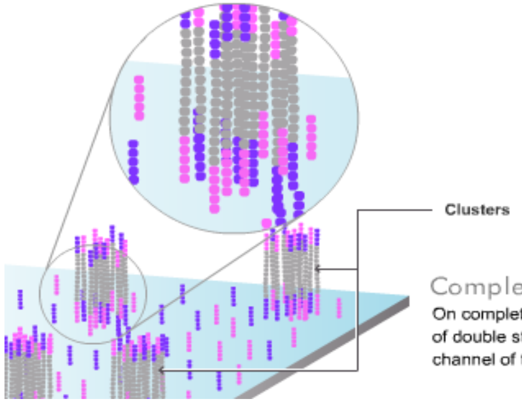






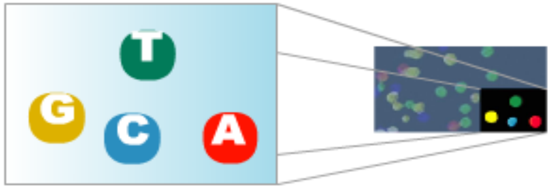
Attach DNA to surface

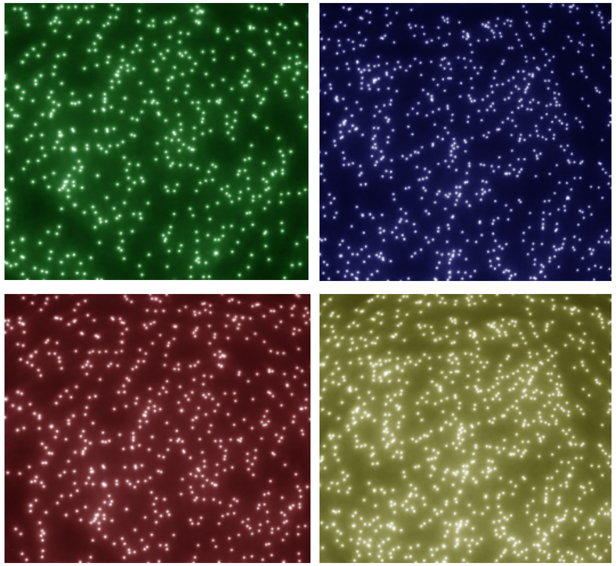
Bind single stranded fragments randomly to the inside surface of the flow cell channels.

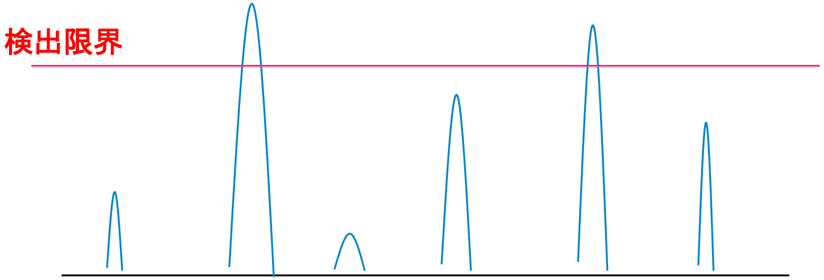


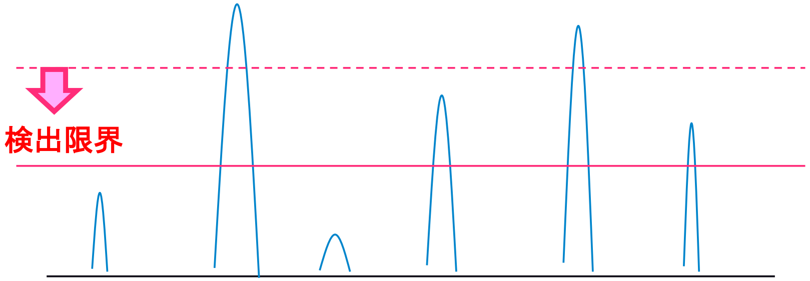
Completion of amplification

On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.





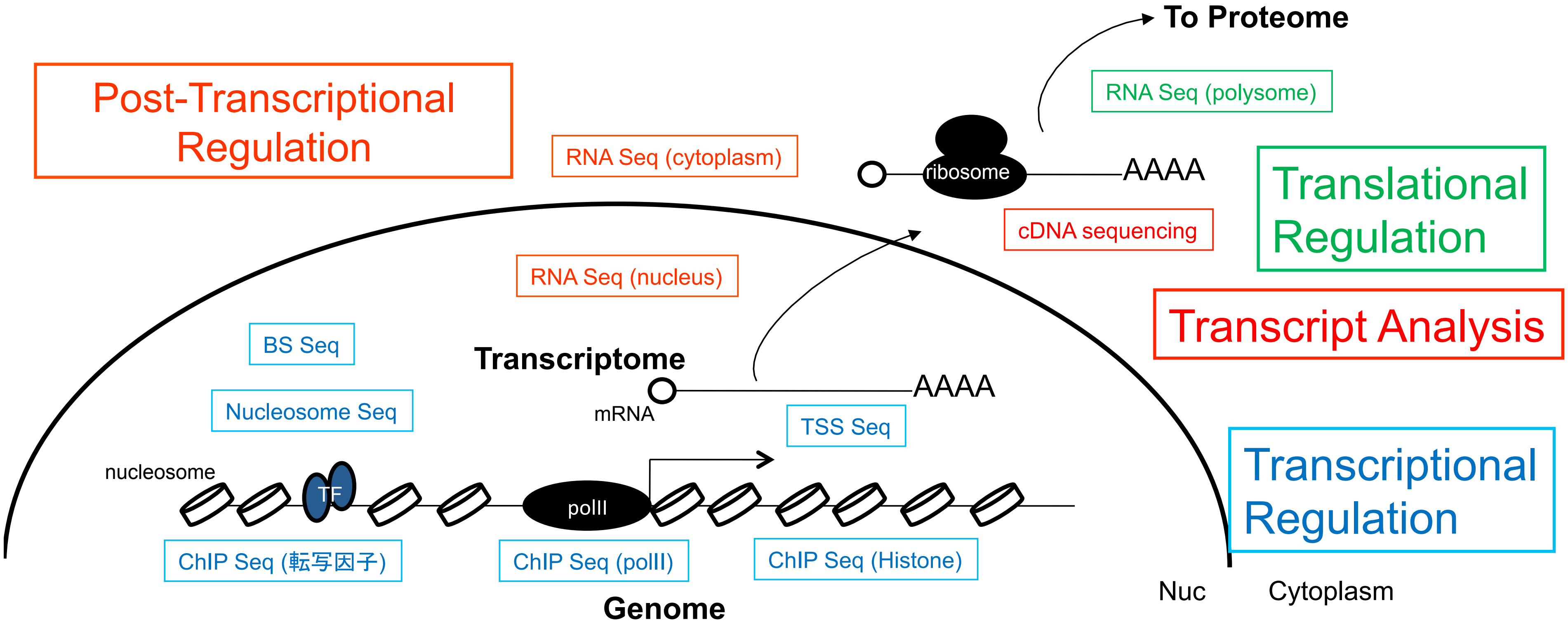




きれいに読めたラン；　機器、画像解析技術の改良

ランごと、スペックごとで読めるDNAが異なる？

“Next Gen” Integrative Genome Analysis



To Proteome

Post-Transcriptional Regulation

RNA Seq (cytoplasm)

RNA Seq (polysome)

cDNA sequencing

Translational Regulation

Transcript Analysis

RNA Seq (nucleus)

BS Seq

Nucleosome Seq

Transcriptome

TSS Seq

Transcriptional Regulation

ChIP Seq (転写因子)

ChIP Seq (polII)

ChIP Seq (Histone)

Genome

Genome-Resequencing

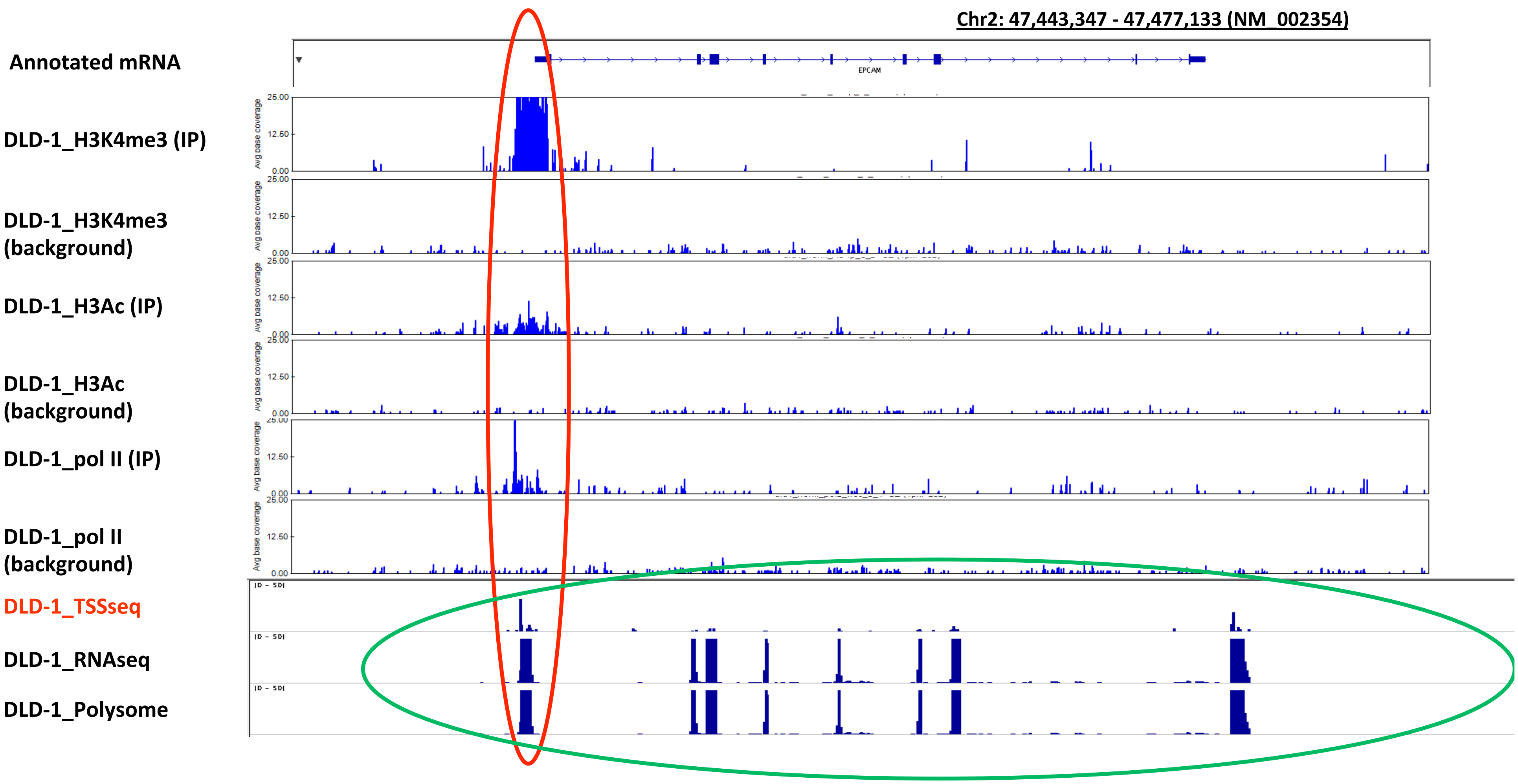
ChIP Seq

RNA Seq

Genome-wide Data using NGS as a Common Platform

次世代シークエンスデータの統合的解析

DLD-1 cell (colon caner)



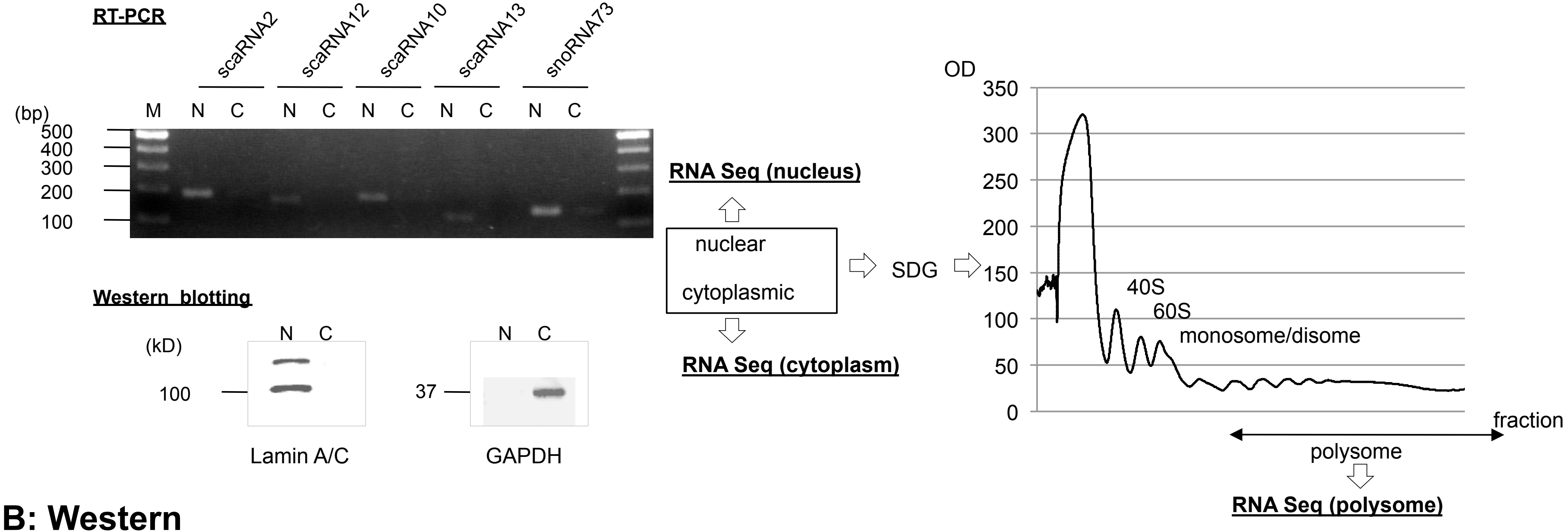
転写制御の網羅的理解へ

Chr2: 47,443,347 - 47,477,133 (NM_002354)

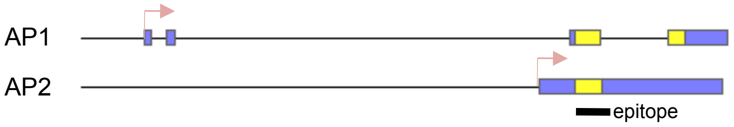
mRNA動態の網羅的理解へ

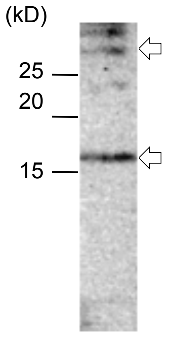
Considering polysome RNA Seq tags for further filtering TSSs

A: 細胞内画分の分画



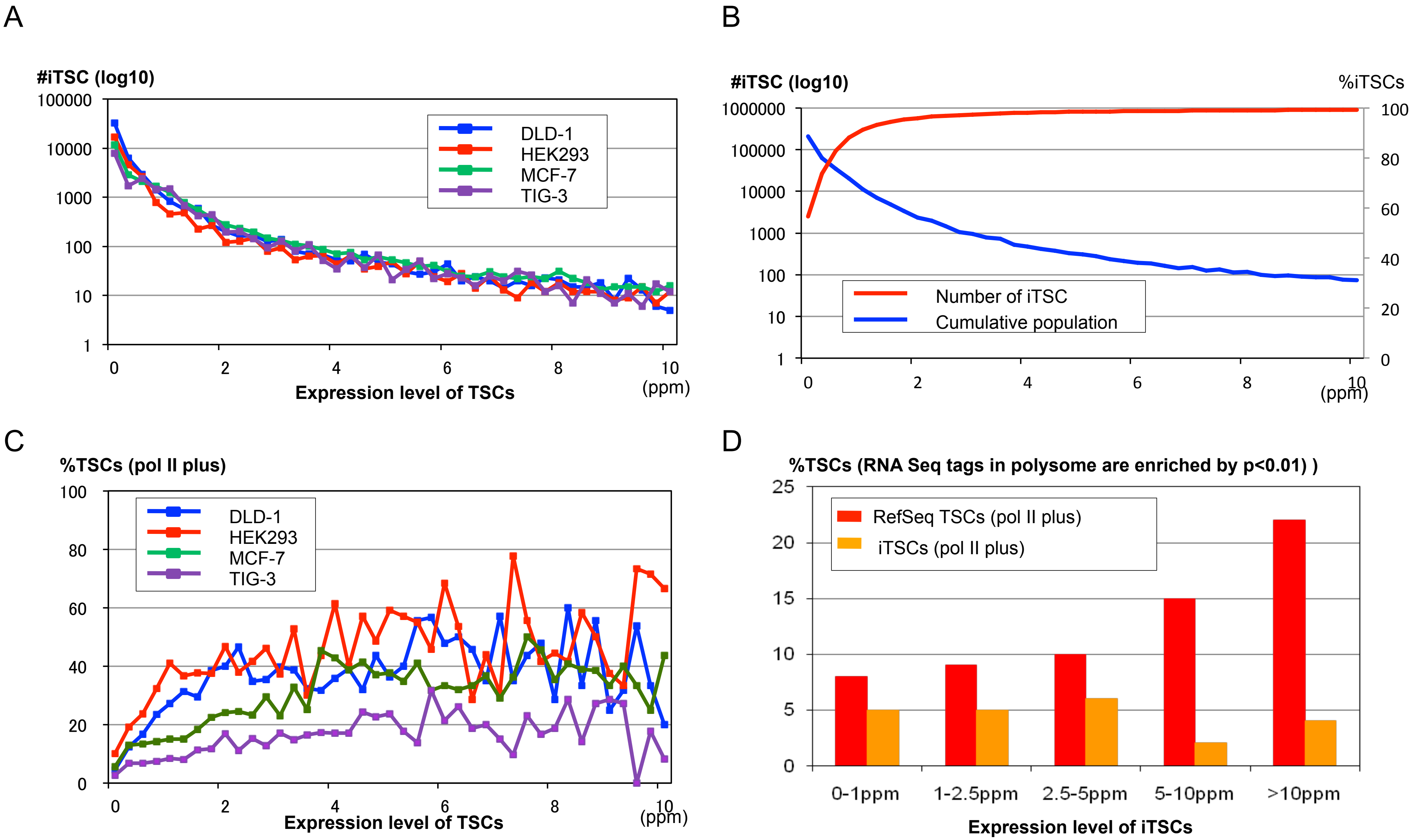
B: Western





NM_018952 (HoxB6)	AP1	AP2
TSS tag	1170ppm	25ppm
pol II binding	bound	bound
Corresponding cDNA	BC014651	X58431 (HIT00195371)*1
Longest ORF in the cDNA	224 aa	140 aa
Expected molecular weight of protein product	25.4 kD	15.2 kD
Translation caveat	no	no
polysome tag enrich (p value)	1e−7	2e−3
RNA Seq tag (polysome)	9ppm	5ppm

Filtration for intergenic TSSs (lncRNAs)



Yamashita et al Genome Res 2011

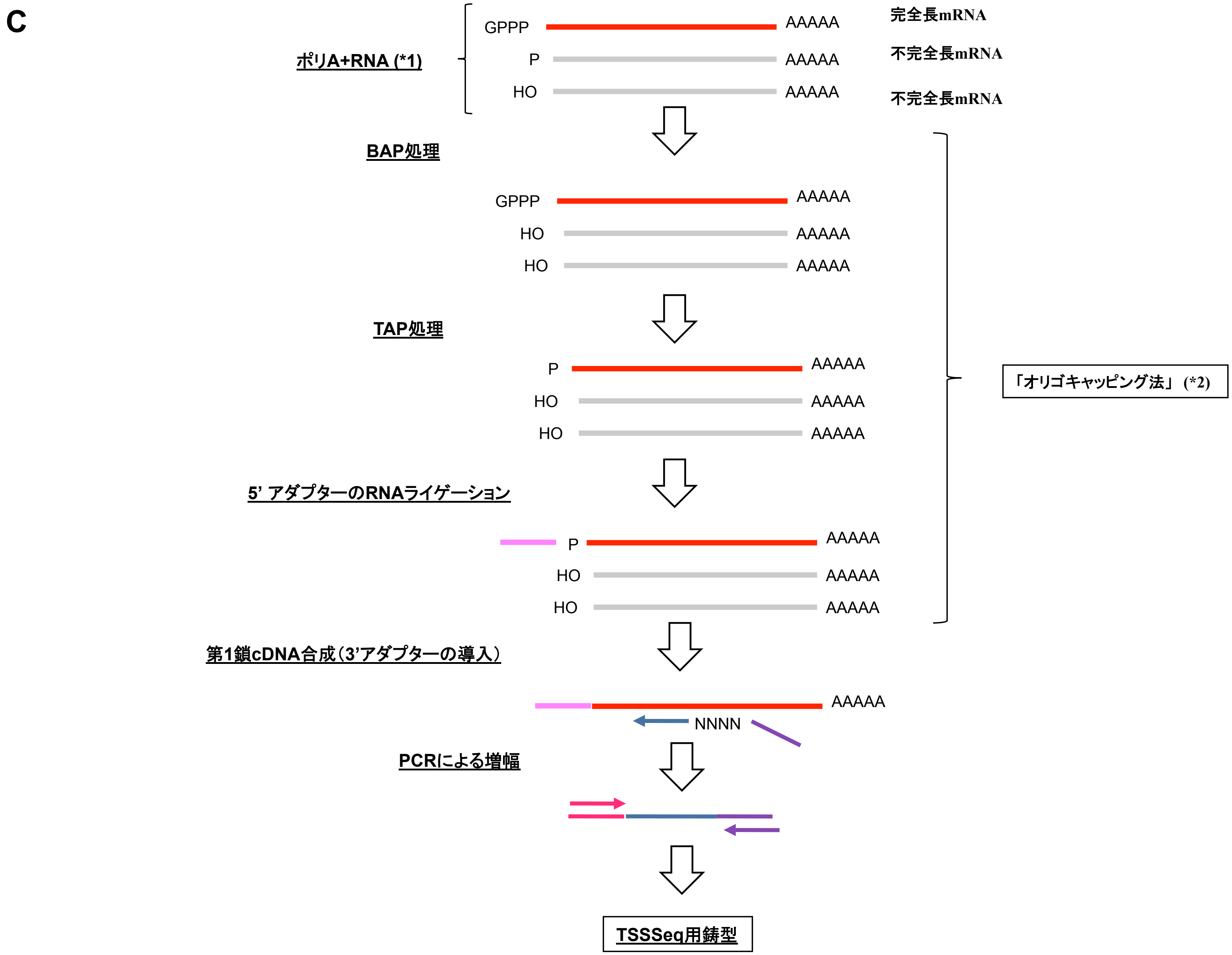

C
ポリA+RNA (*1)
GPPP
AAAAA
完全長mRNA
P
AAAAA
不完全長mRNA
HO
AAAAA
不完全長mRNA
BAP処理
GPPP
AAAAA
HO
AAAAA
HO
AAAAA
TAP処理
P
AAAAA
HO
AAAAA
HO
AAAAA
「オリゴキャッピング法」 (*2)
5' アダプターのRNAライゲーション
P
AAAAA
HO
AAAAA
HO
AAAAA
第1鎖cDNA合成（3'アダプターの導入）
AAAAA
NNNN
PCRによる増幅
TSSSeq用鋳型

図1C

TSS Seq (DLD-1; the GRHL3 gene region)

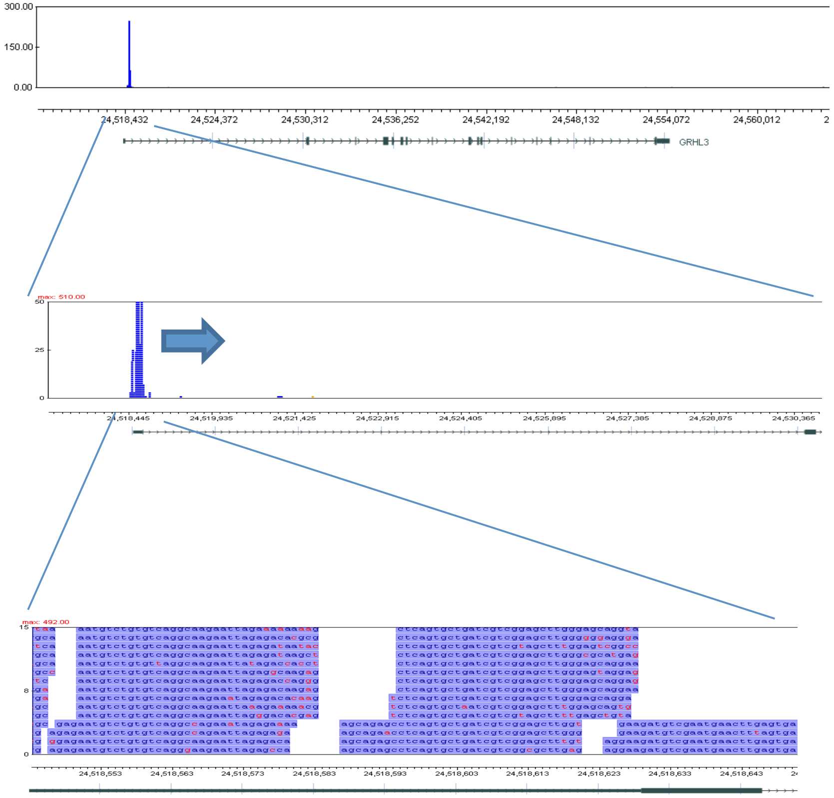
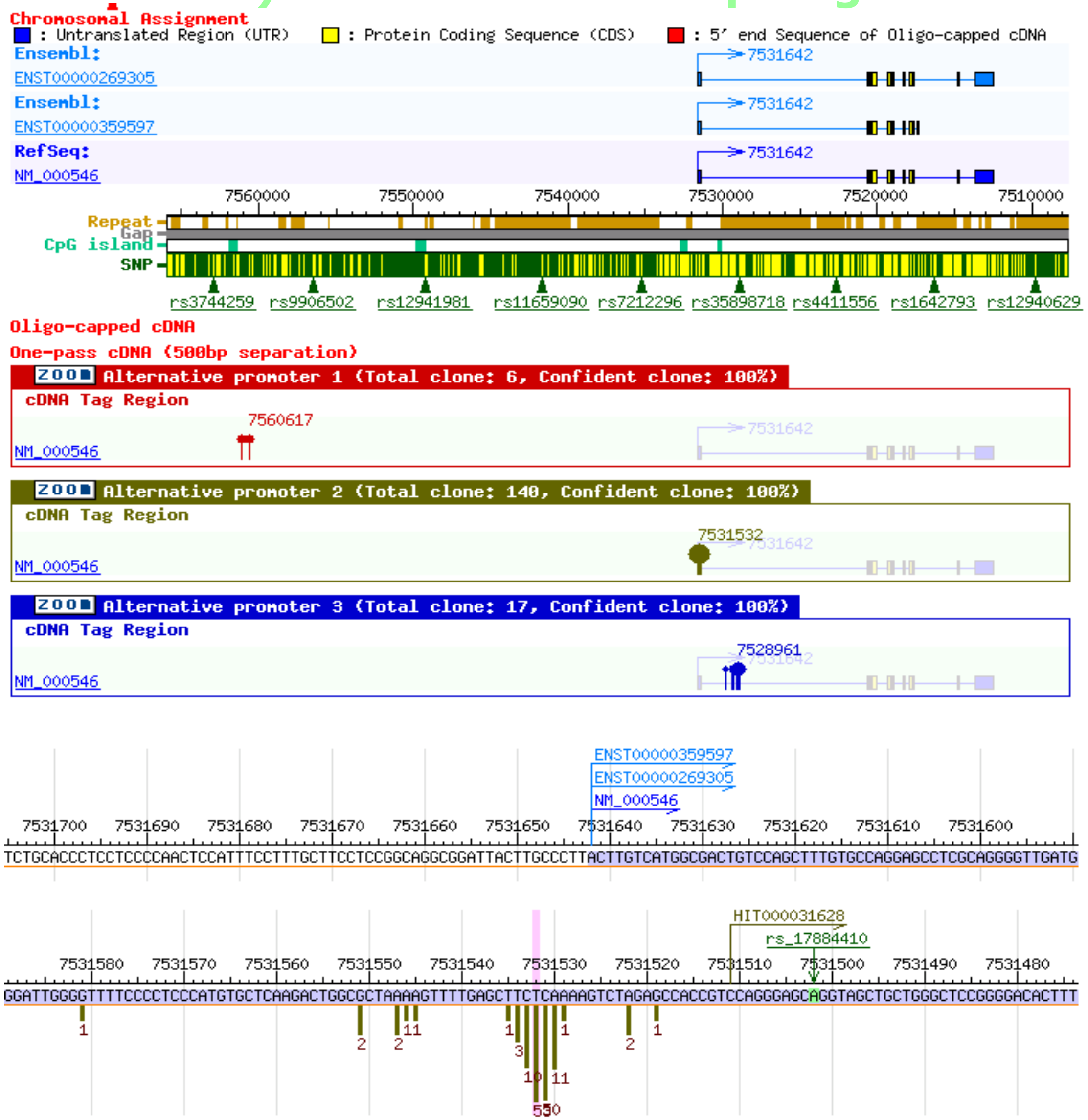


図2C

(Alternative) Promoters of the p53 gene in MCF7 cells



6 TSS-tags

140 TSS-tags= Previously reported "TSS"

17 TSS-tags

Major promoter correctly detected

Detection of Minor promoters

“Intergenic” TSSs

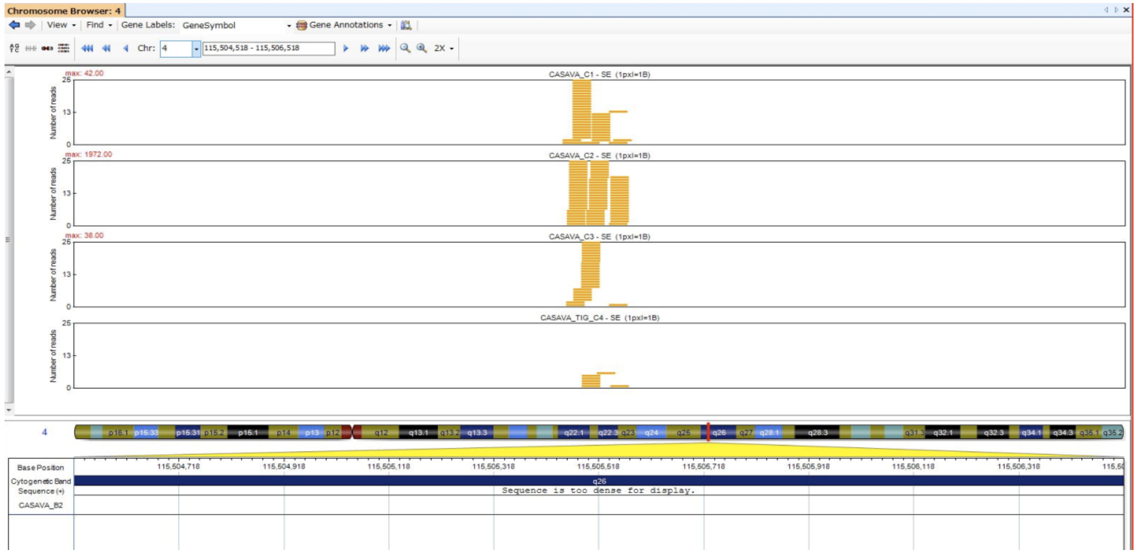
Analysis of full-length cDNAs and TSS Seq

Control

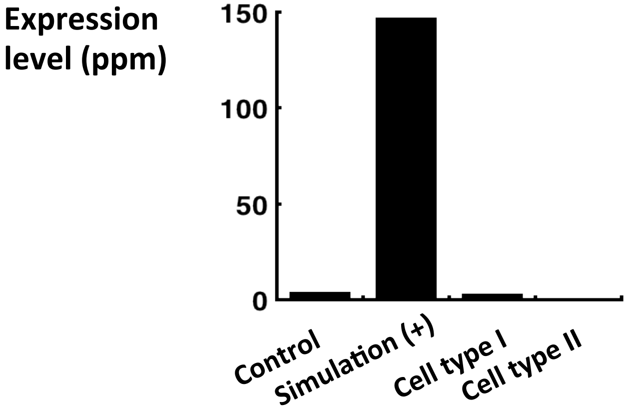
Simulation (+)

Cell type I

Cell type II

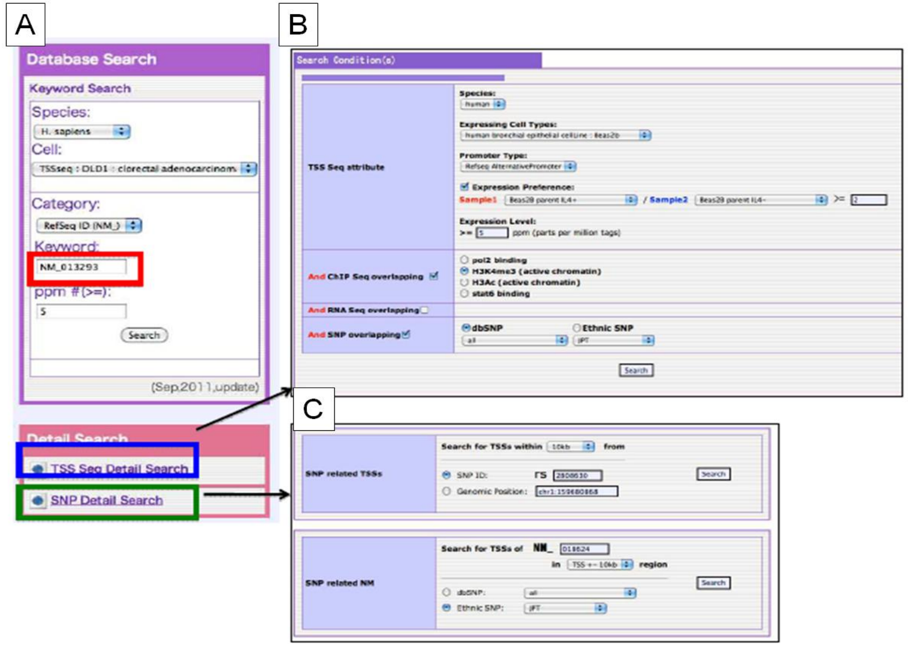


Validation analysis by qRT-PCR



Biological relevance of long ncRNAs still remains mostly unknown…

DataBase of Transcriptional Start Sites (DBTSS@http://dbtss.hgc.jp/)



Species: H. sapiens
Cell: TSSseq : DLD1 : clorectal adenocarcinoma
Category: RefSeq ID (NM_)
Keyword: NM_013293
ppm #(>=): 5

Species: human
Expressing Cell Types: human broechial epithelial celline : Beas2b
Promoter Type: Refseq AlternativePromoter
☑ Expression Preference:
Sample1 Beas2B parent IL4+ / Sample2 Beas2B parent IL4- >= 2
Expression Level:
>= 5 ppm (parts per million tags)

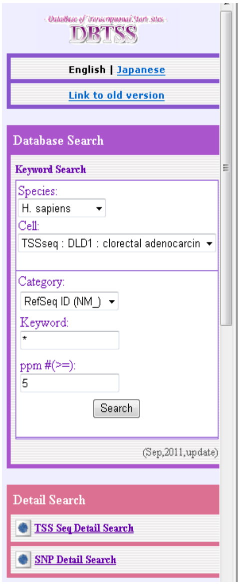
And ChIP Seq overlapping ☑
○ pol2 binding
● H3K4me3 (active chromatin)
○ H3Ac (active chromatin)
○ stat6 binding

And RNA Seq overlapping ☐

And SNP overlapping ☑
● dbSNP all
○ Ethnic SNP JPT

Search for TSSs within 10kb from
● SNP ID: rs 2808630
○ Genomic Position: chr1:159680868

Search for TSSs of NM_ 018624
in TSS +- 10kb region
○ dbSNP: all
● Ethnic SNP: JPT



Species: H. sapiens
Cell: TSSseq : DLD1 : clorectal adenocarcin
Category: RefSeq ID (NM_)
Keyword: *
ppm #(>=): 5

Release 8.0 Updated (September 15 2011)
Based on UCSC hg19, mm9
We recommend to use the Internet Explorer 6.0 or later for visiting our database.

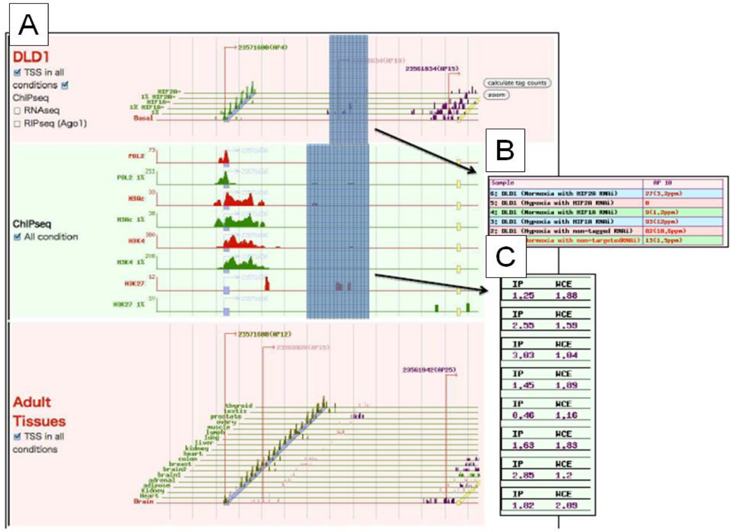
About this Database

DBTSS: Database of Transcriptional Start Sites

Current version is based on UCSC hg19, mm9

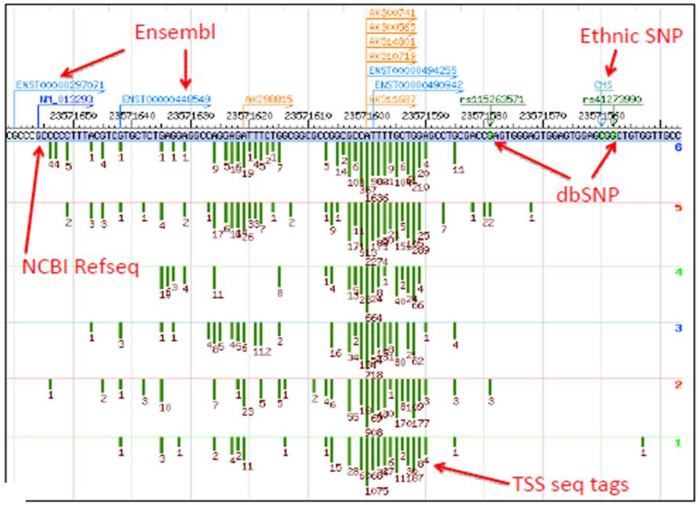
ABSTRACT | About this Database

To support transcriptional regulation studies, we have constructed the DBTSS (DataBase of Transcriptional Start Sites), which represnts exact positions of transcriptional start sites (TSSs) in the genome based on our unique experimentally validated TSS sequencing method, TSS Seq. Here we included new TSS data, so that a major part of human adult and embryonic tissues are covered. DBTSS now contains 491 million TSS tag sequences for collected from a total of xxx tissue and cell cultures. We also integrated our newly generated RNA-seq data of subcellular- fractionated RNAs and ChIP Seq data of histone modifications, RNA polymerase II and several transcriptional regulatory factors in cultured cell lines. We also included recently accumulating external epigenomic data, such as chromatin map of the ENCODE project. We further associated those TSS information with public and original SNV data, in order to indentify single nucleotide variations (SNVs) in the regulatory regions. These data can be browsed in our new viewer which also supports versatile search conditions of users. We believe new DBTSS is helpful to understand biological consequences of the massively identified TSSs and identify human genetic valuations which are associated with disordered transcriptional regulations.

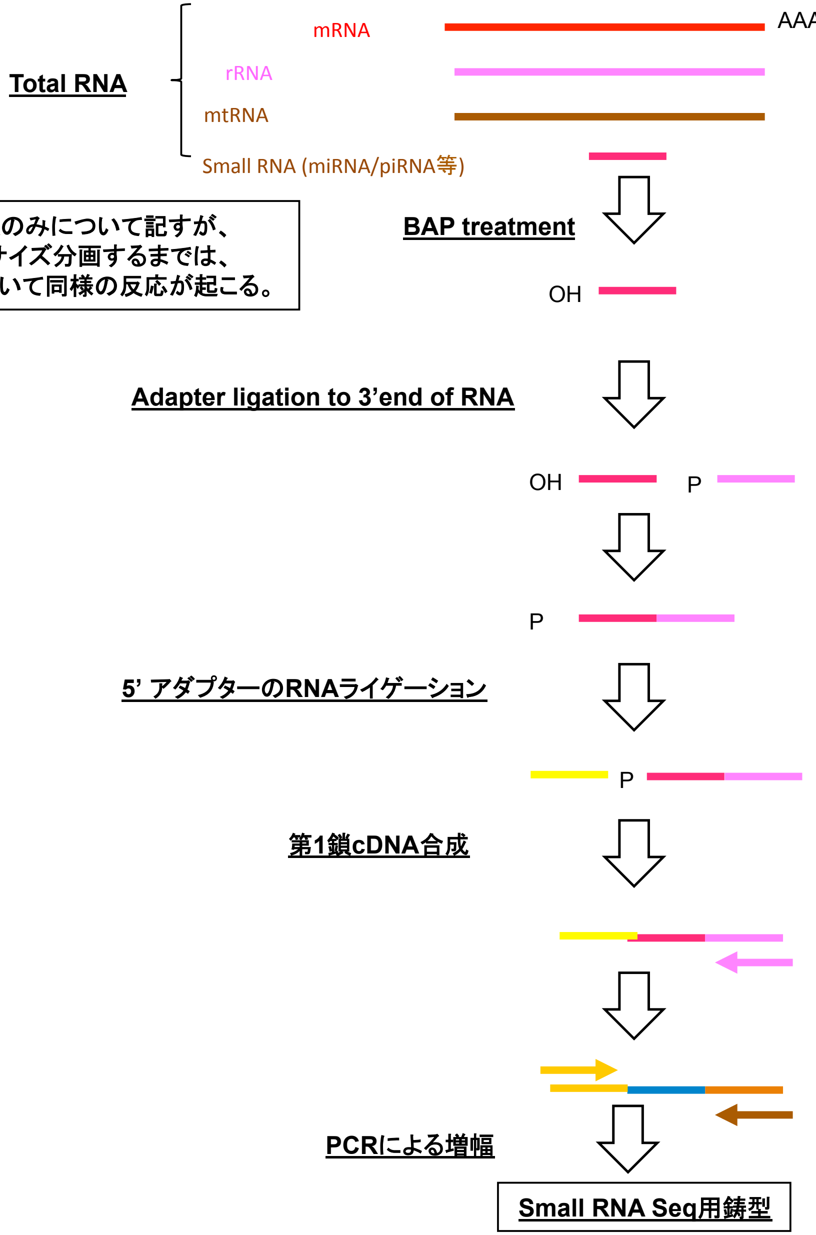


Sample	HP 1B
6: DLD1 (Normoxia with HIF2A RNAi)	27(3.2ppm)
5: DLD1 (Hypoxia with HIF2A RNAi)	0
4: DLD1 (Normoxia with HIF1A RNAi)	9(1.2ppm)
3: DLD1 (Hypoxia with HIF1A RNAi)	93(12ppm)
2: DLD1 (Hypoxia with non-tagged RNAi)	82(10.6ppm)
ormoxia with non-targetedRNAi)	13(1.5ppm)

IP	WCE
1.25	1.88
2.55	1.59
3.03	1.04
1.45	1.89
0.46	1.16
1.63	1.83
2.85	1.2
1.82	2.09



（おもに）miRNA の解析



Total RNA

mRNA AAAAA

rRNA

mtRNA

Small RNA (miRNA/piRNA等)

BAP treatment

OH

Adapter ligation to 3’end of RNA

OH P

P

5’ アダプターのRNAライゲーション

P

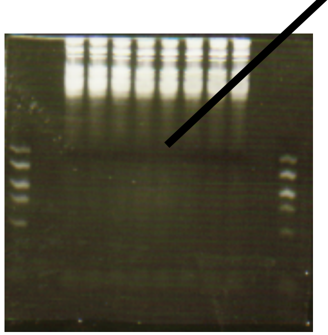
第1鎖cDNA合成

PCRによる増幅

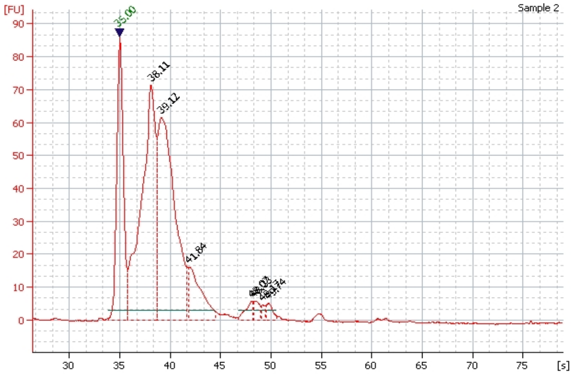
Small RNA Seq用鋳型

※図はsmall RNAのみについて記すが、最後のステップでサイズ分画するまでは、すべてのRNAについて同様の反応が起こる。

	Takara Protocol	Illumina protocol (v1.5)
Total RNA input	100ug	1ug
Size selection	Need	No need



約18 nt~30 nt分画のSmall RNAを単離



small RNA Seq (DLD-1; the MIMAT0004584 gene region)

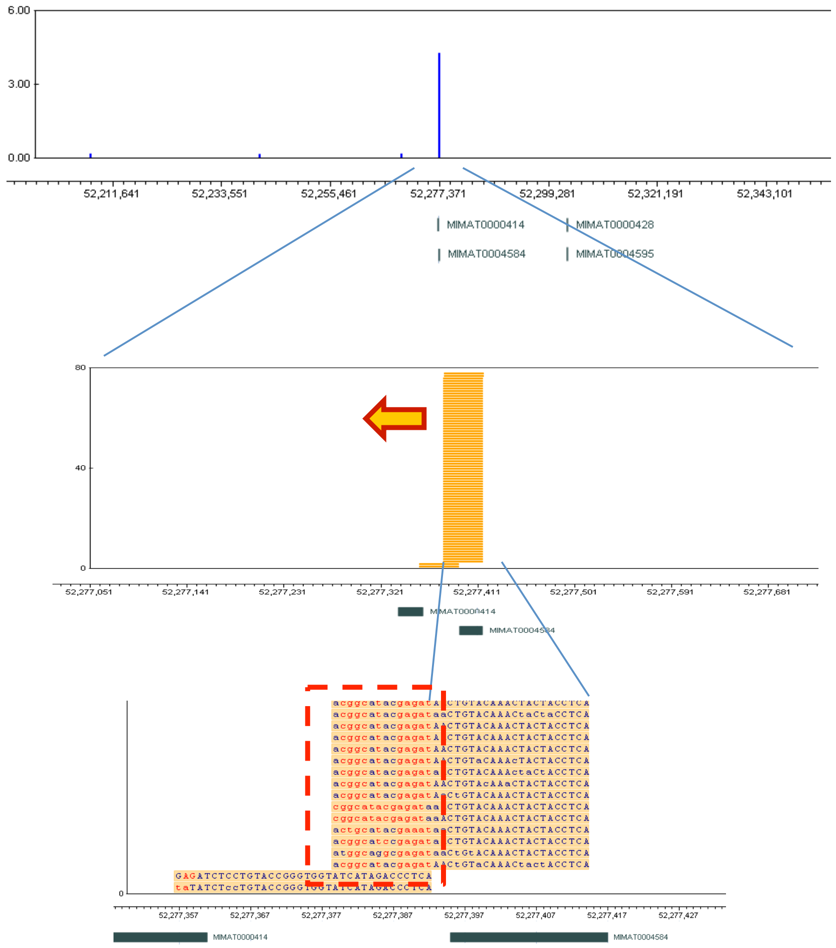
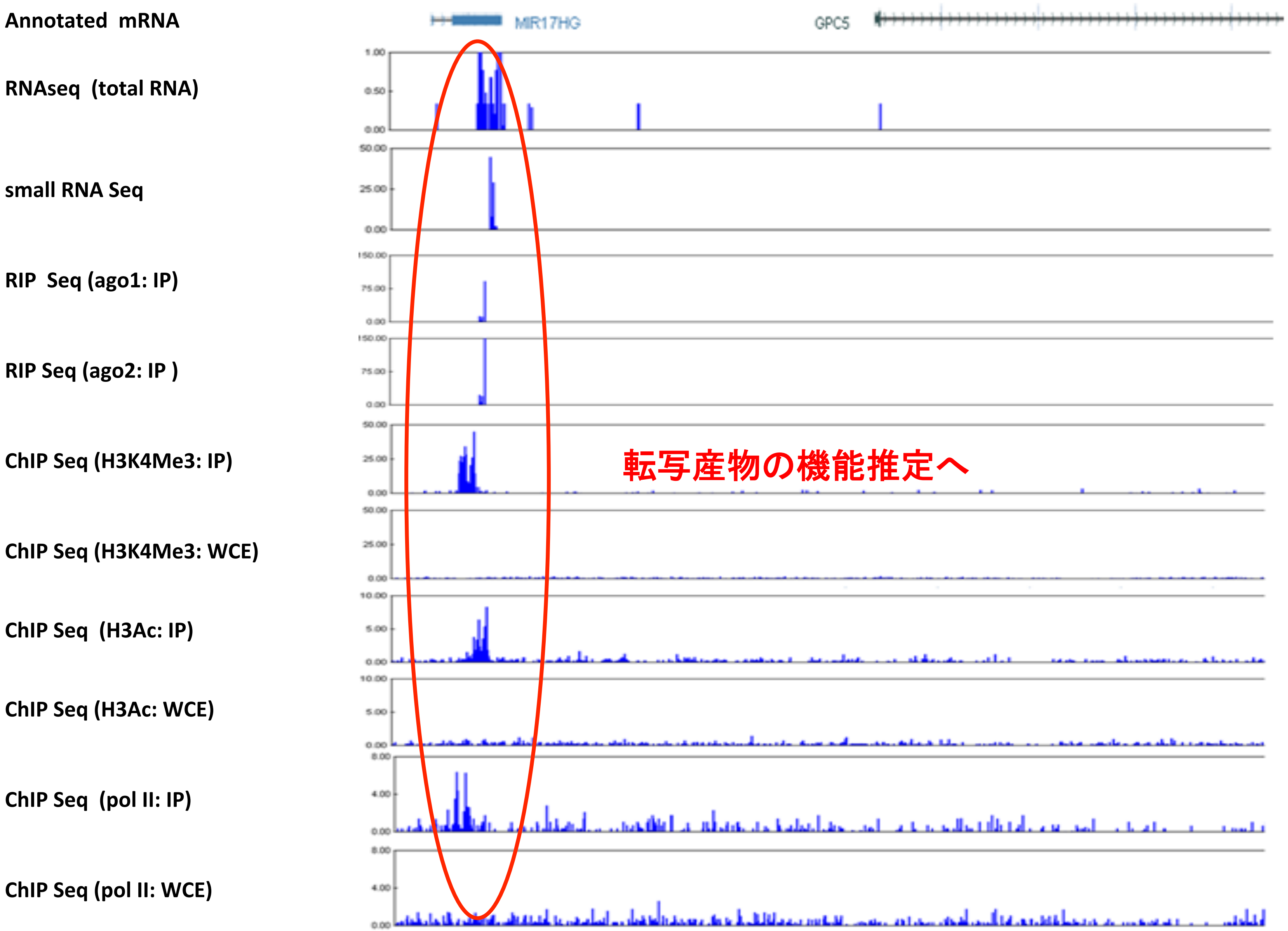


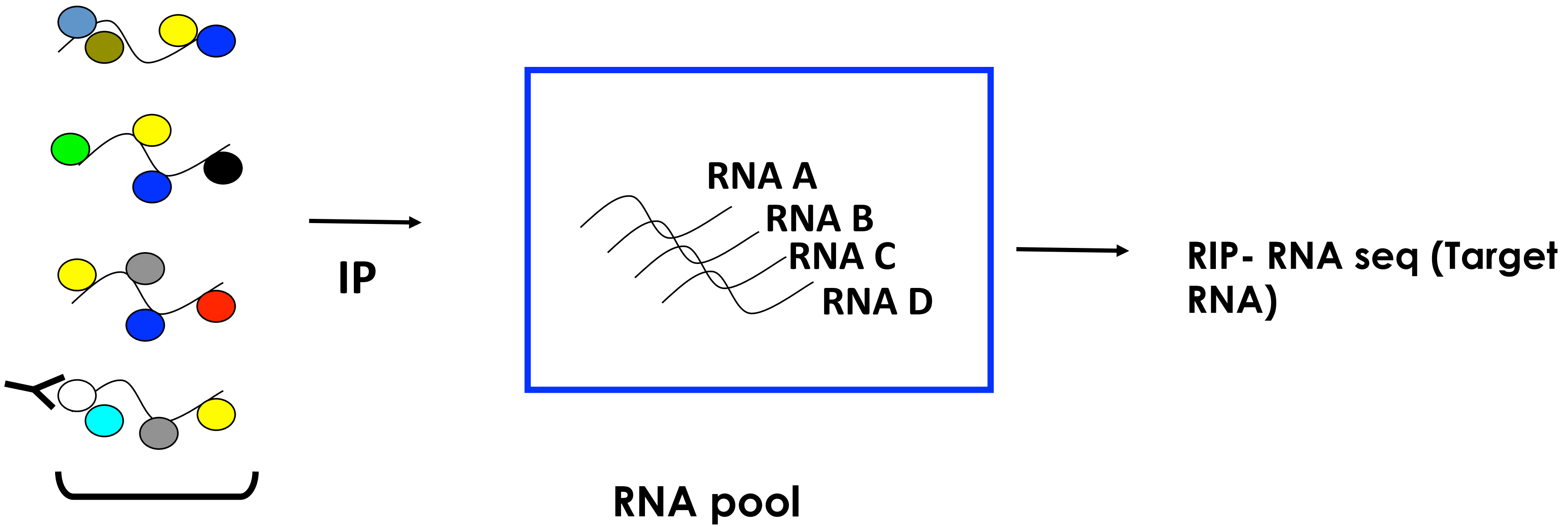
図2B

B **The MIR17HG gene region (DLD-1 cells)**



転写産物の機能推定へ

Schematic diagram of RIP(RNA immunoprecipitation) -Seq



IP

RNA A

RNA B

RNA C

RNA D

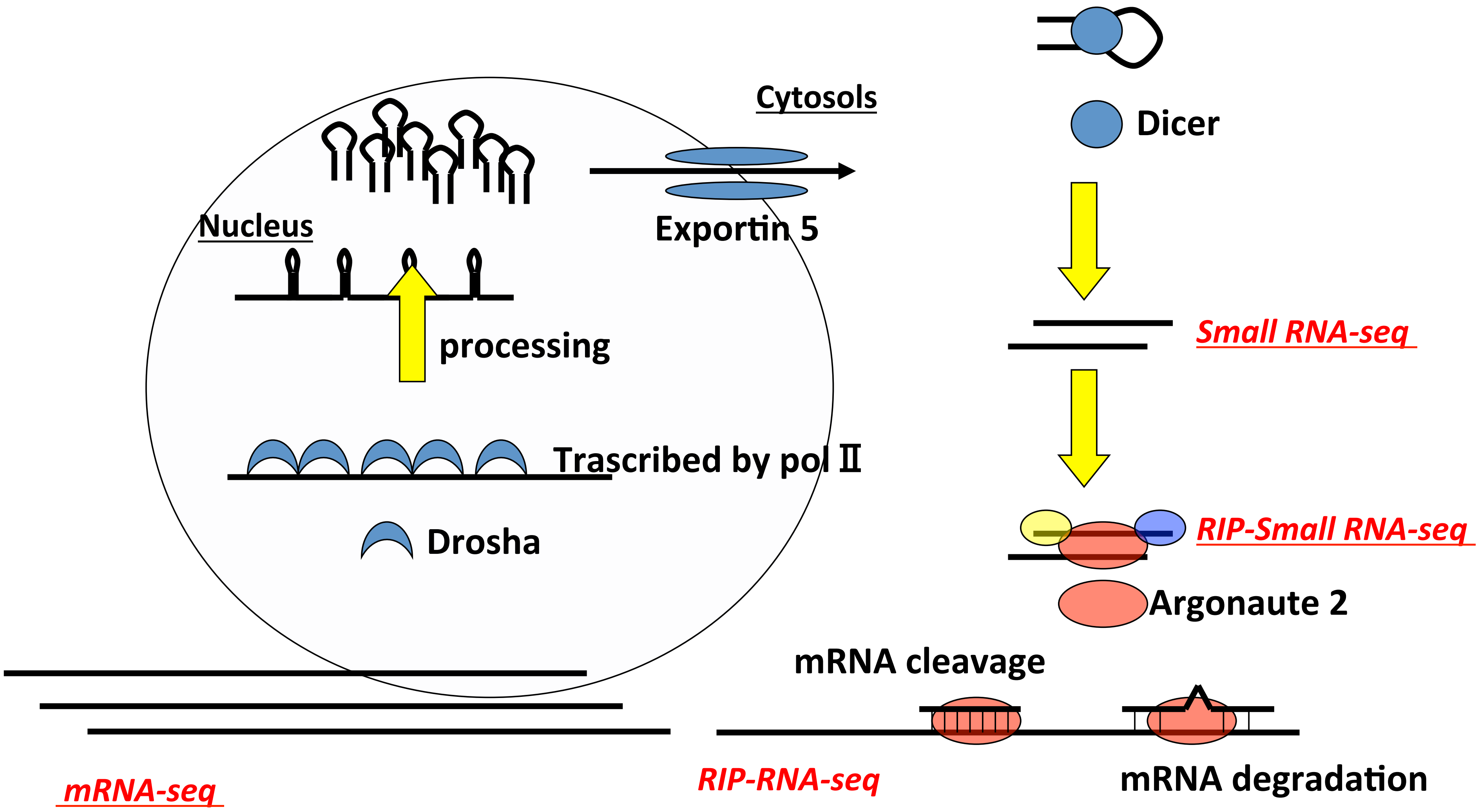
RIP- RNA seq (Target RNA)

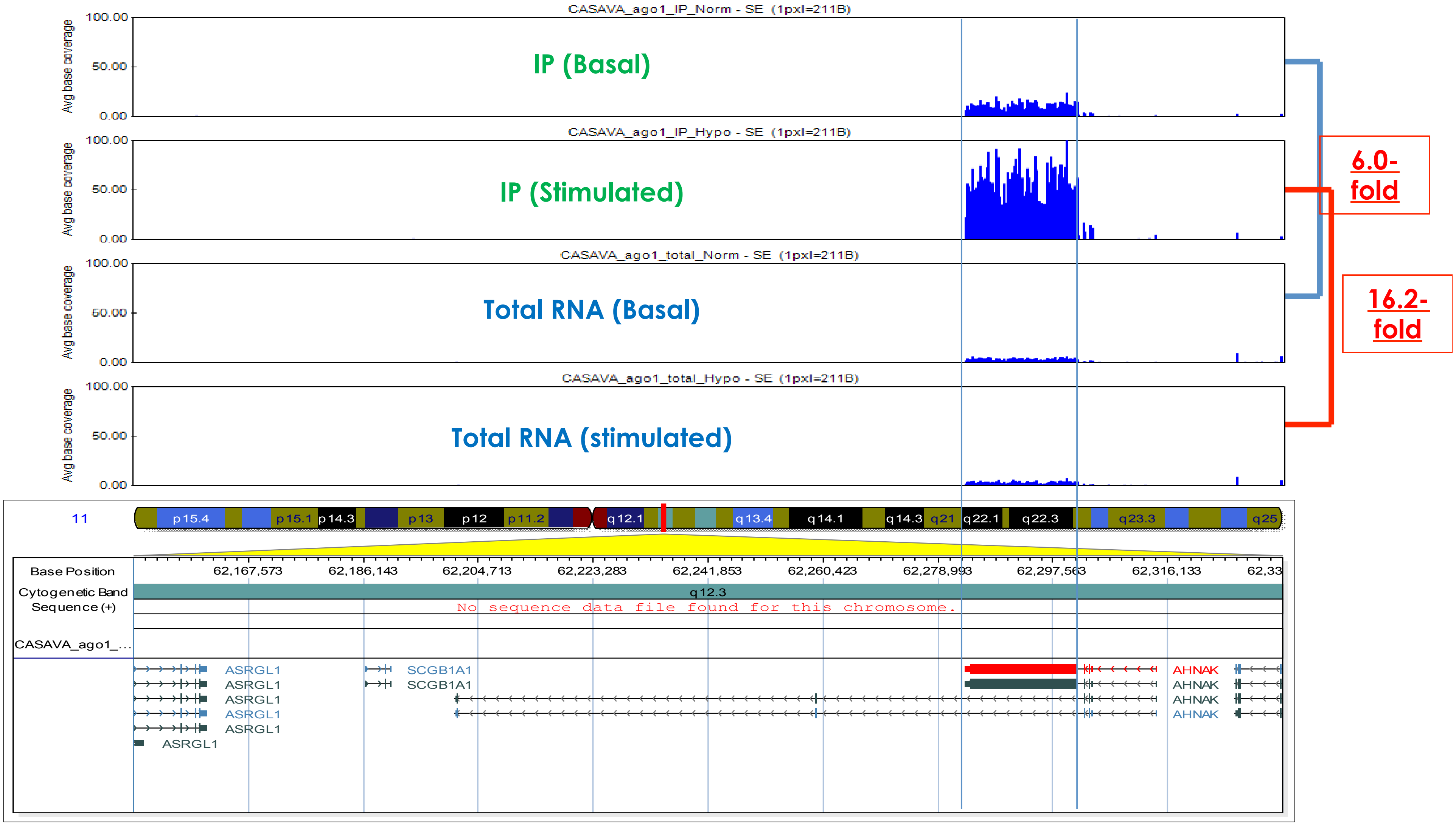
RNA pool

● RNA binding protein

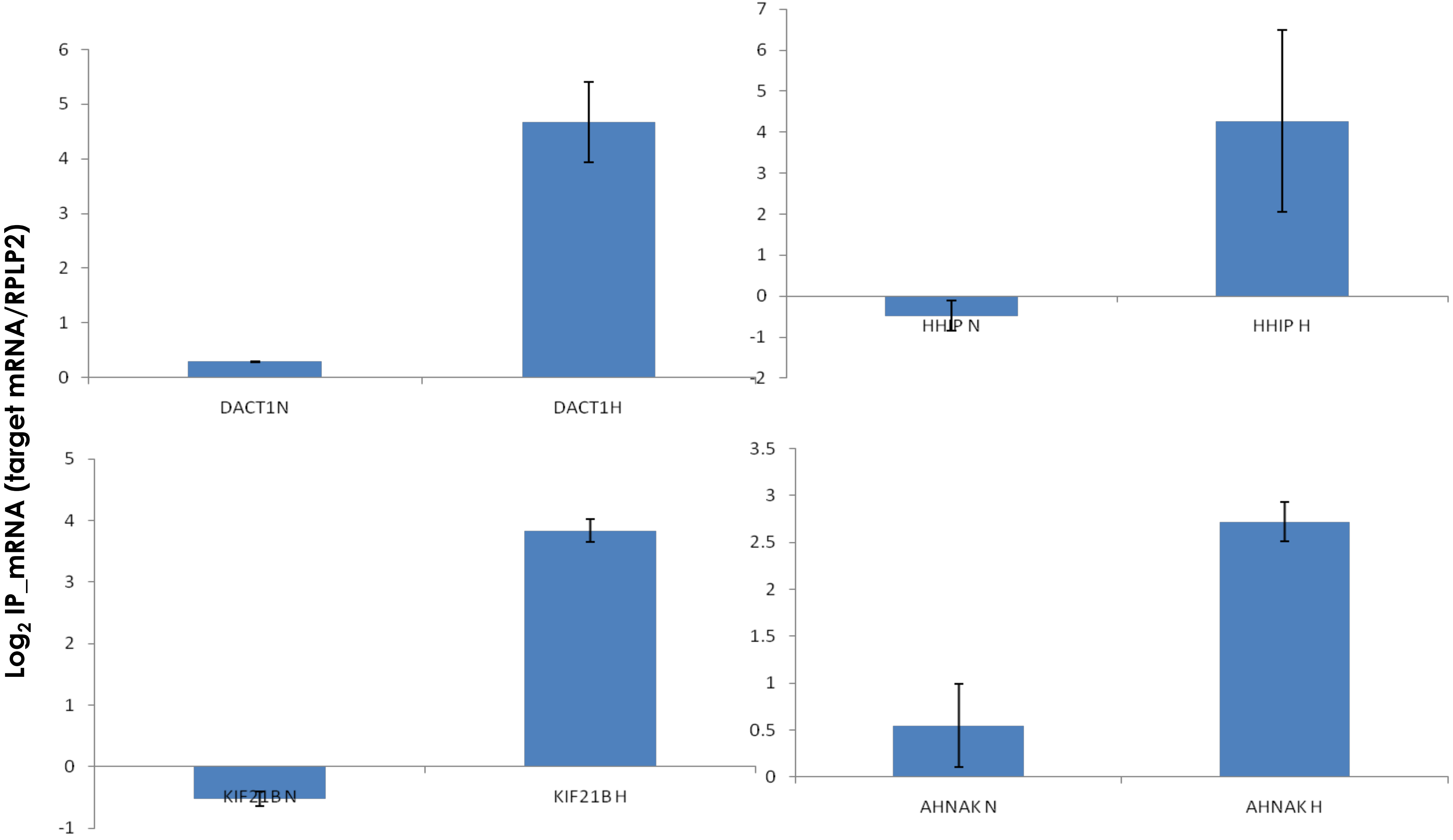
〜 Target RNA

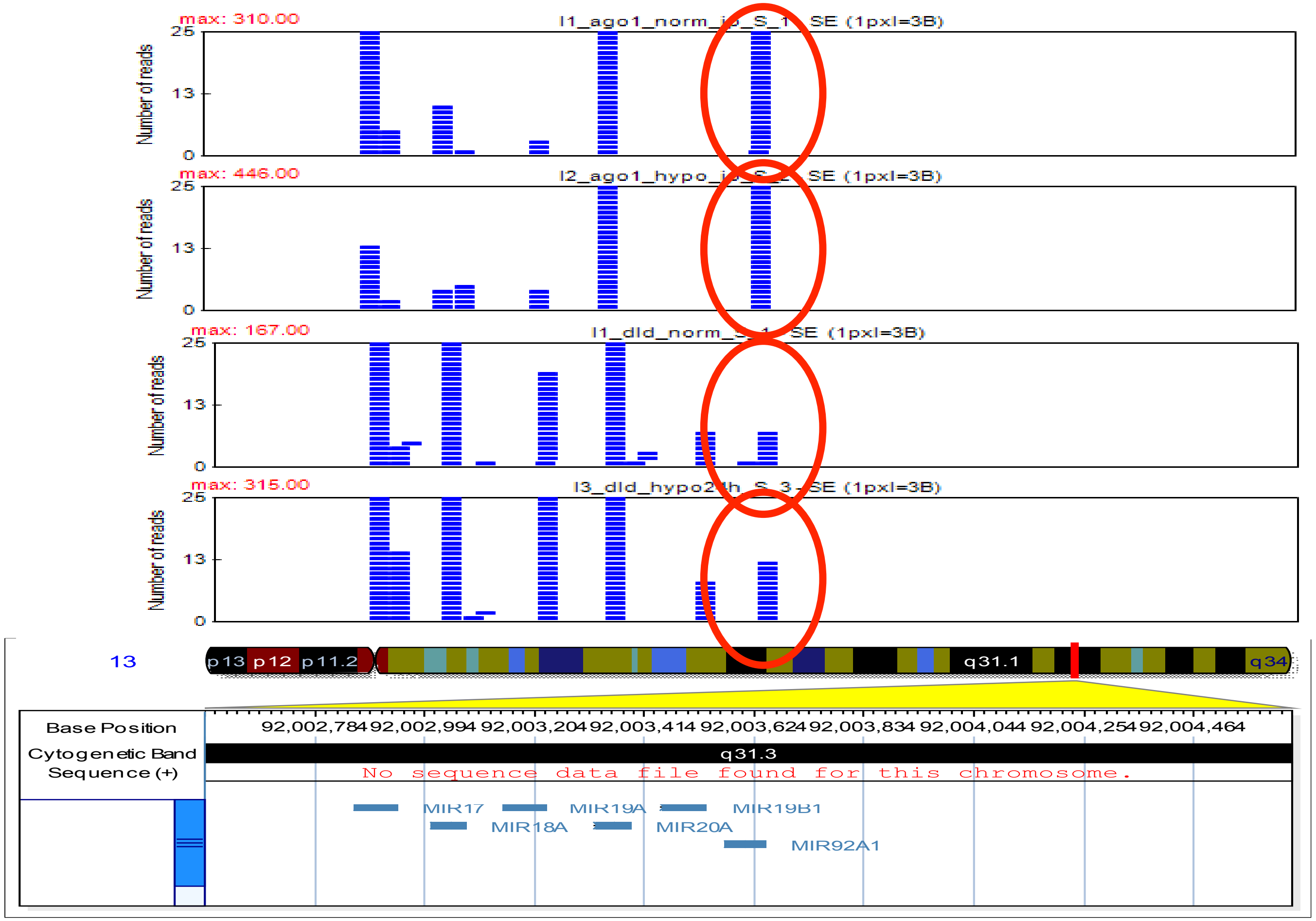
Schematic diagram of biogenesis of microRNAs and post-transcriptional silencing of target mRNA




CASAVA_ago1_IP_Norm - SE (1pxl=211B)
IP (Basal)
CASAVA_ago1_IP_Hypo - SE (1pxl=211B)
IP (Stimulated)
CASAVA_ago1_total_Norm - SE (1pxl=211B)
Total RNA (Basal)
CASAVA_ago1_total_Hypo - SE (1pxl=211B)
Total RNA (stimulated)
Avg base coverage
100.00
50.00
0.00
6.0-fold
16.2-fold
11
p15.4
p15.1
p14.3
p13
p12
p11.2
q12.1
q13.4
q14.1
q14.3
q21
q22.1
q22.3
q23.3
q25
Base Position
62,167,573
62,186,143
62,204,713
62,223,283
62,241,853
62,260,423
62,278,993
62,297,563
62,316,133
62,33
Cytogenetic Band
q12.3
Sequence (+)
No sequence data file found for this chromosome.
CASAVA_ago1_...
ASRGL1
SCGB1A1
AHNAK

Validation By Real-time PCR 1
(N>=3)





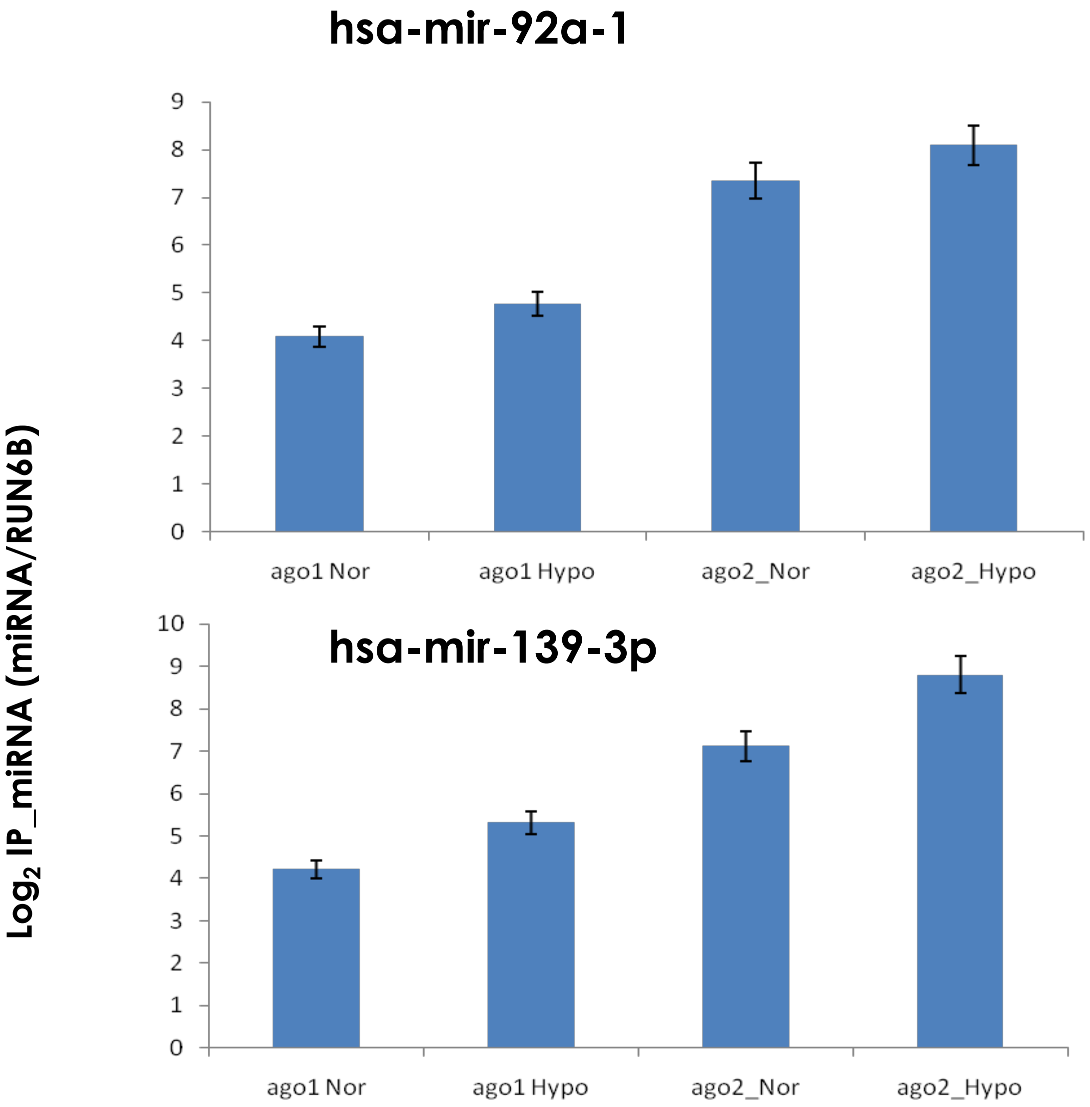
IP (Basal)

IP (Stimulated)

Total RNA (Basal)

Total RNA (stimulated)

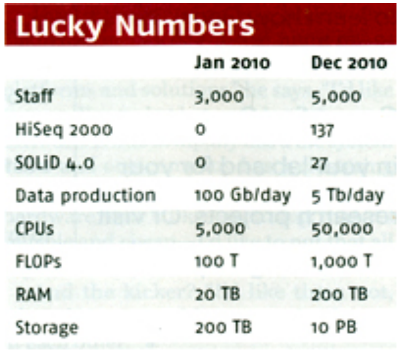
Validation By Real-time PCR (N=3)



多検体化へ
＝自動化の普及

"Sample Prep Automation" in Big/Small Sequence Centers

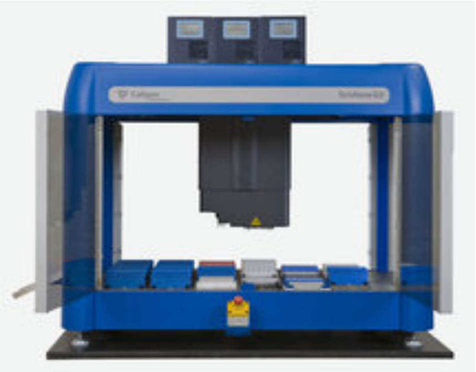
Beijin Genome Institute (China: >100 HiSeq)



Lucky Numbers

	Jan 2010	Dec 2010
Staff	3,000	5,000
HiSeq 2000	0	137
SOLiD 4.0	0	27
Data production	100 Gb/day	5 Tb/day
CPUs	5,000	50,000
FLOPs	100 T	1,000 T
RAM	20 TB	200 TB
Storage	200 TB	10 PB

"Sciclone" (Caliper)

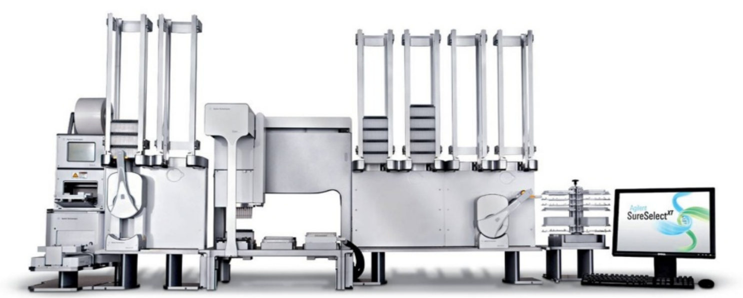


BROAD Institute (US: >50 HiSeq)

96MTP in 3 days

AgilentXT

+Exon-capture automation



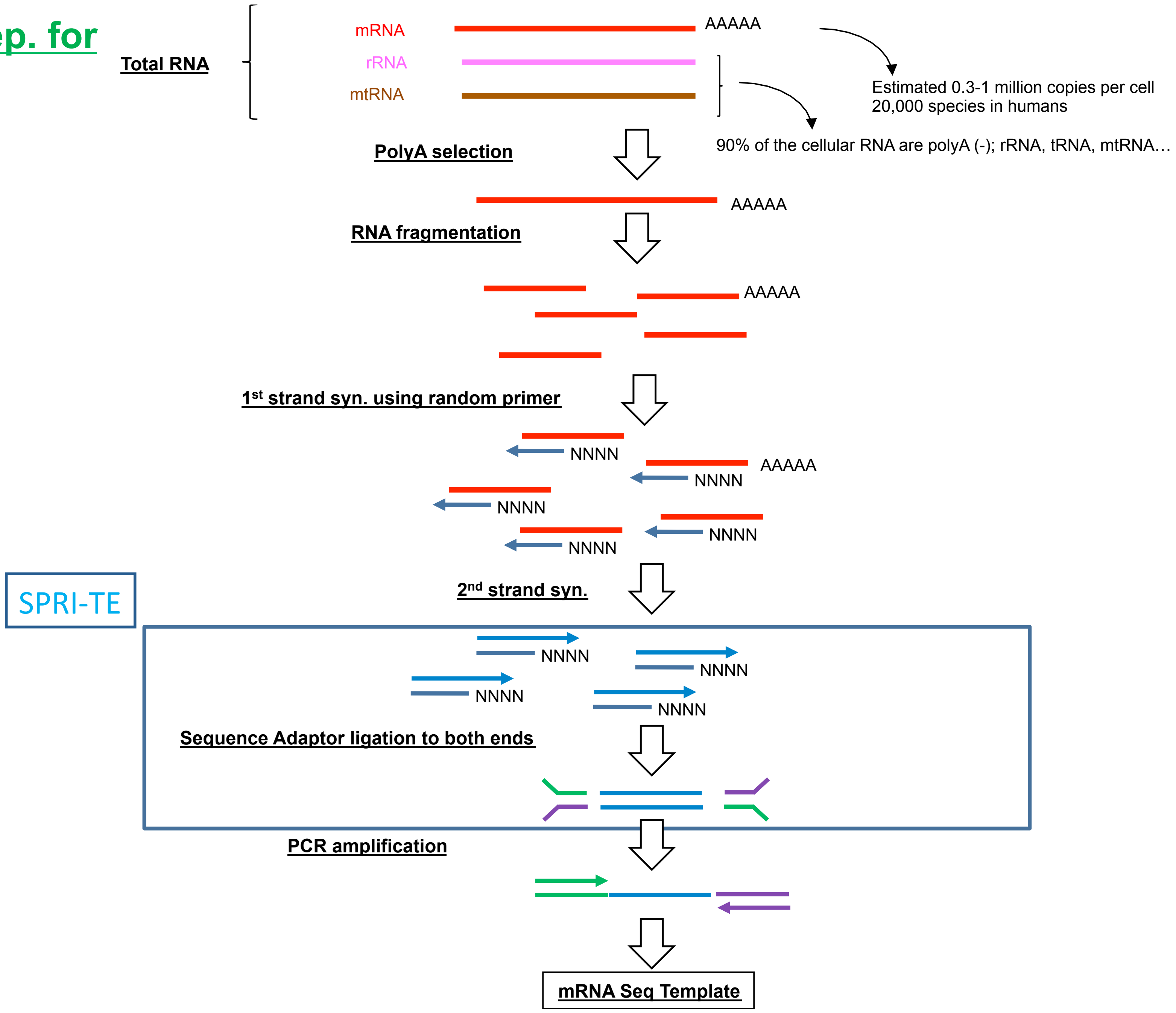
Kashiwa, UT(Japan: 1 HiSeq + 3GAIIx)

12 sample Double strand DNA/cDNA

Adaptor ligation

Size fractionation (gel-free)

Template Prep. for RNA Seq



Total RNA

Estimated 0.3-1 million copies per cell
20,000 species in humans

90% of the cellular RNA are polyA (-); rRNA, tRNA, mtRNA…

PolyA selection

RNA fragmentation

1st strand syn. using random primer

2nd strand syn.

SPRI-TE

Sequence Adaptor ligation to both ends

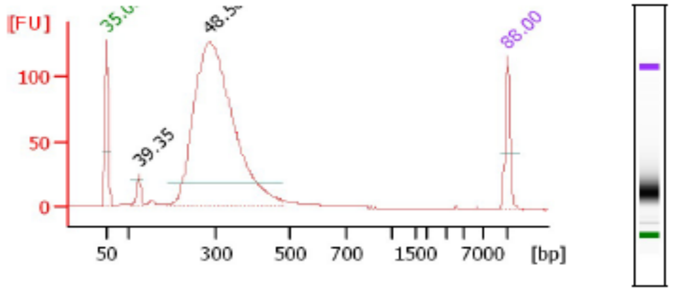
PCR amplification

mRNA Seq Template

“Acceptable” Sample Prep

Lot#: 023554

Input: 10 ng



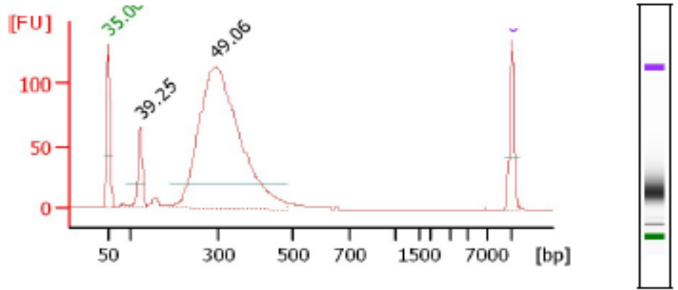
Overall Results for sample 3 : input 10ng lot 022433

Number of peaks found: 2

Peak table for sample 3 : input 10ng lot 022433

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	124	1.55	19.0	
3	284	61.40	327.7	

IP: 2 ng



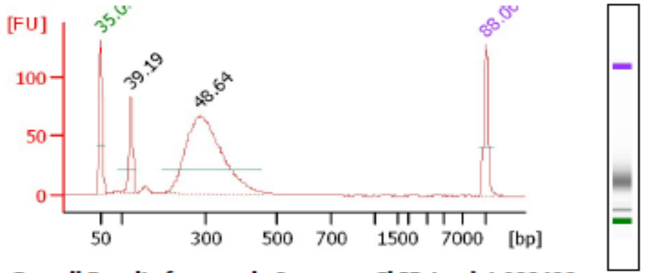
Overall Results for sample 7 : ChIP 2ng lot 022433

Number of peaks found: 2

Peak table for sample 7 : ChIP 2ng lot 022433

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	122	3.96	49.1	
3	292	52.87	274.1	

IP: 1 ng



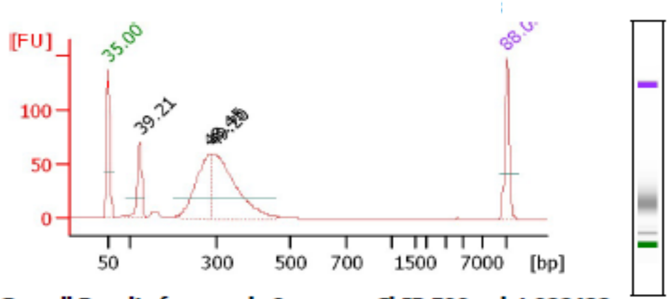
Overall Results for sample 8 : ChIP 1ng lot 022433

Number of peaks found: 2

Peak table for sample 8 : ChIP 1ng lot 022433

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	121	4.83	60.3	
3	285	28.22	150.1	

IP: 500 pg



Overall Results for sample 9 : ChIP 500pg lot 022433

Number of peaks found: 3

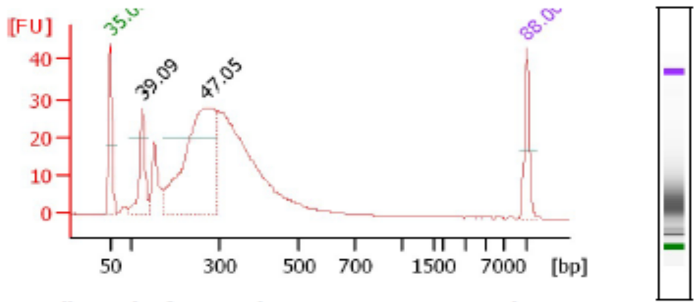
Peak table for sample 9 : ChIP 500pg lot 022433

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	122	4.04	50.3	
3	282	9.75	52.5	
4	295	14.62	75.2	

“Un-Acceptable” Sample Prep

Lot#: 022433

Input: 10 ng



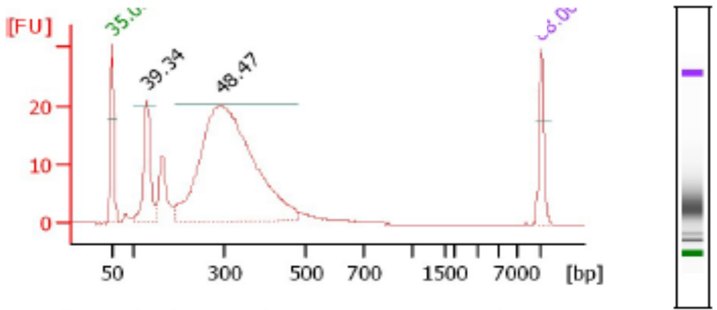
Overall Results for sample 6 : input 10ng lot023554

Number of peaks found: 2

Peak table for sample 6 : input 10ng lot023554

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	126	6.11	73.2	
3	268	27.66	156.6	

IP: 2 ng



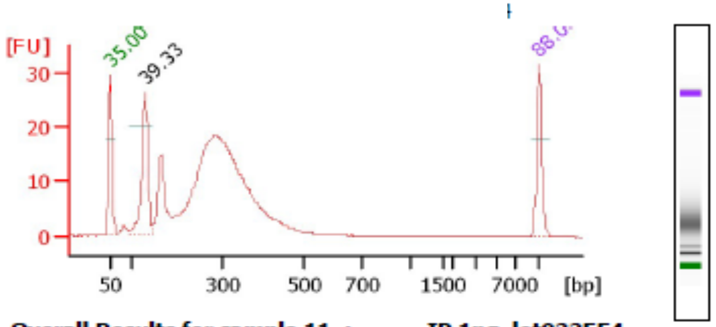
Overall Results for sample 10 : IP 2ng lot023554

Number of peaks found: 2

Peak table for sample 10 : IP 2ng lot023554

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	131	7.06	81.7	
3	293	45.23	233.9	

IP: 1 ng



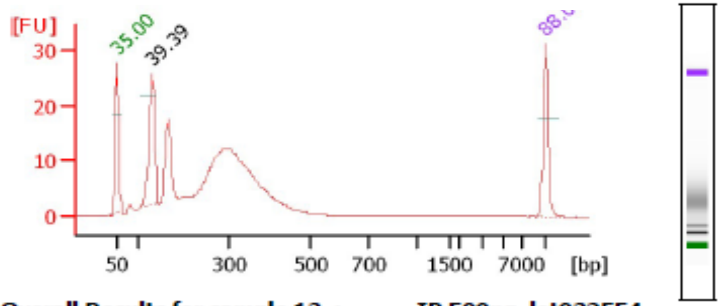
Overall Results for sample 11 : IP 1ng lot023554

Number of peaks found: 1

Peak table for sample 11 : IP 1ng lot023554

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	131	8.26	95.8	

IP: 500 pg



Overall Results for sample 12 : IP 500ng lot023554

Number of peaks found: 1

Peak table for sample 12 : IP 500ng lot023554

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
2	132	6.80	78.3	

Sequence outcomes of SPRI-TE samples

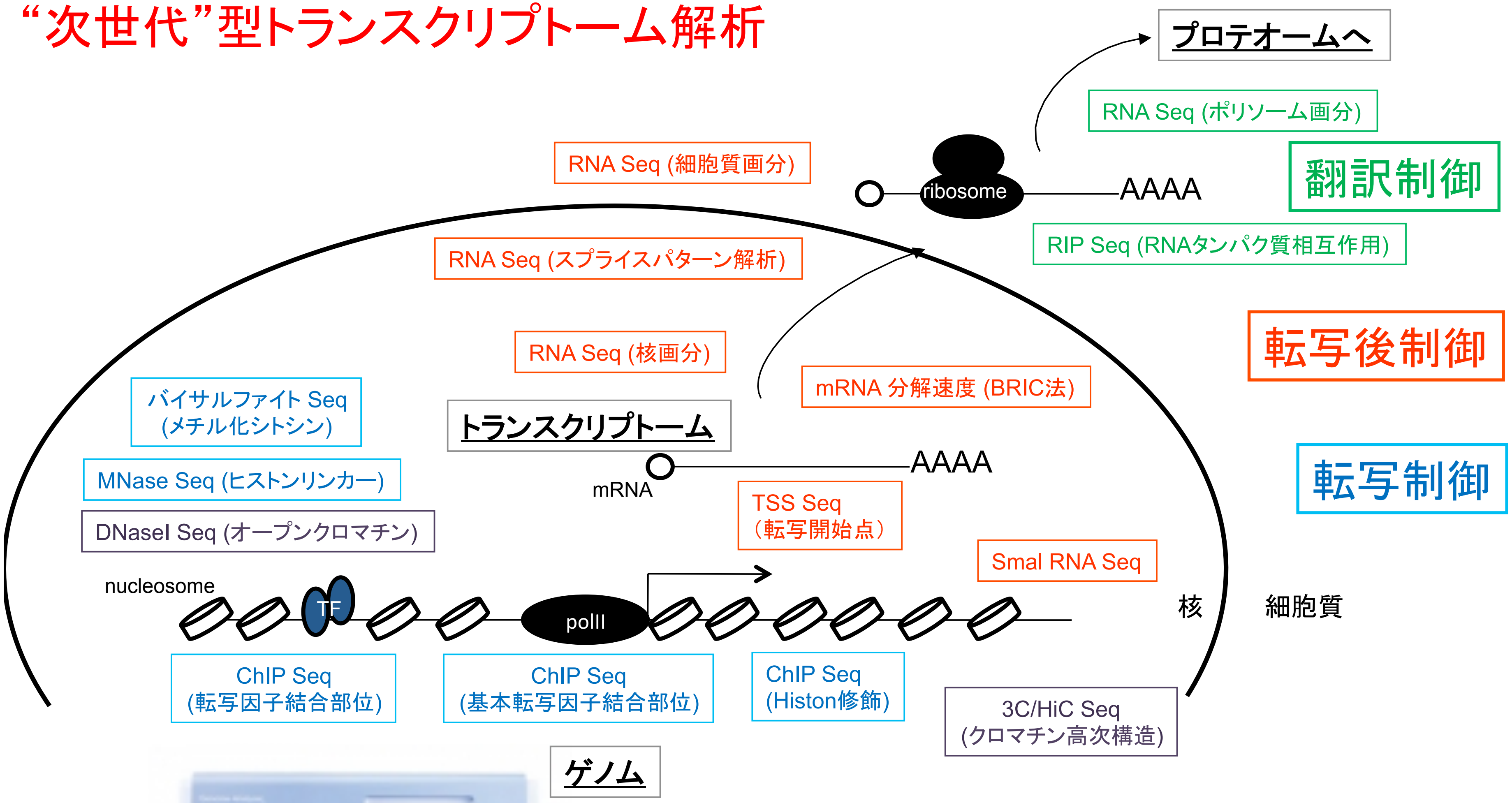
■Favorable samples

Sample ID	# tota l read	primer dimer	dimer (%)
sampleA	28,502,937	559,185	2.0%
sampleB	28,635,753	660,799	2.3%

■Unfavorable samples

サンプル名	総リード数	primerDimer	全タグ中の割合
sampleC	42,781,509	8,797,067	20.6%
sampleD	35,953,985	4,553,405	12.7%

“次世代”型トランスクリプトーム解析



プロテオームへ

RNA Seq (ポリソーム画分)

RNA Seq (細胞質画分)

ribosome

AAAA

翻訳制御

RIP Seq (RNAタンパク質相互作用)

RNA Seq (スプライスパターン解析)

転写後制御

RNA Seq (核画分)

mRNA 分解速度 (BRIC法)

バイサルファイト Seq
(メチル化シトシン)

トランスクリプトーム

MNase Seq (ヒストンリンカー)

mRNA

AAAA

転写制御

TSS Seq
（転写開始点）

DNaseI Seq (オープンクロマチン)

Smal RNA Seq

nucleosome

TF

polII

核

細胞質

ChIP Seq
(転写因子結合部位)

ChIP Seq
(基本転写因子結合部位)

ChIP Seq
(Histon修飾)

3C/HiC Seq
(クロマチン高次構造)

ゲノム

共通検出器としての次世代シークエンサー

まとめ

Sequencing Design（メインプラットフォームはGAIIx; HiSeqはゲノムリシークエンス用）

>50M reads/lane (>80% data mapped)

Single end read (3 days)

36 bp

TSS Seq: 1 lane

ChIP/RNA IP Seq: 1 lane (WCE/IP)

RNA Seq: 1 lane

Small RNA Seq: 1 lane

おそらくx4ぐらいのindexingは可能

Pair end read (10 days)

76 bp

De novo用RNA Seq: 1 lane

cDNA shotgun: 1000 clone/lane;
(see Reginald et al PLoS One, 2010)

Towards medical appication

Disease

Clinical trial

Cell/Animal model

??

Mal-function of genes

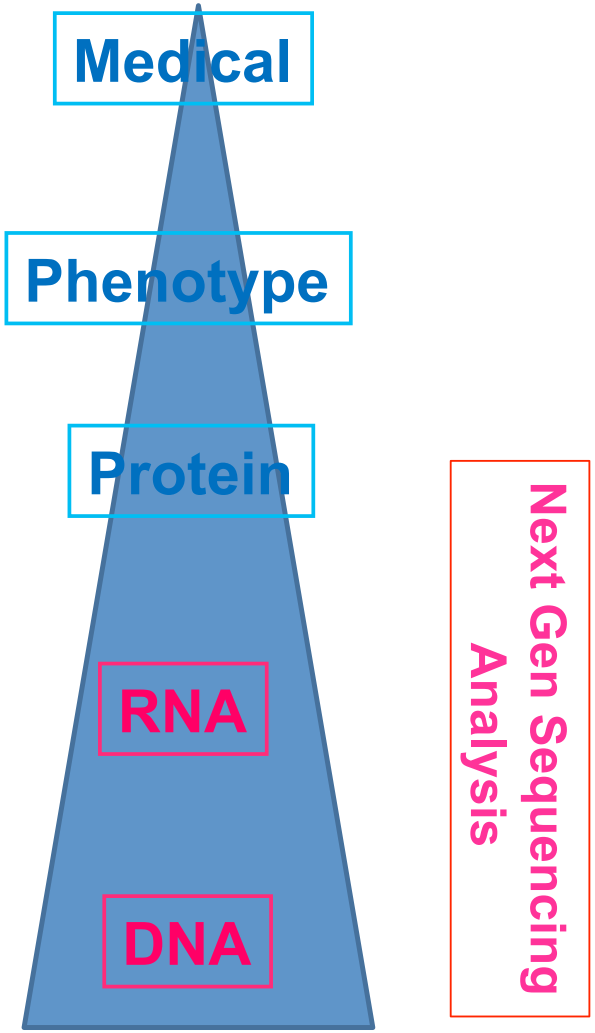
??

Expression disorder

RNA Seq, ChIP Seq

DNA mutation

Exome, Whole genome resequencing



Medical

Phenotype

Protein

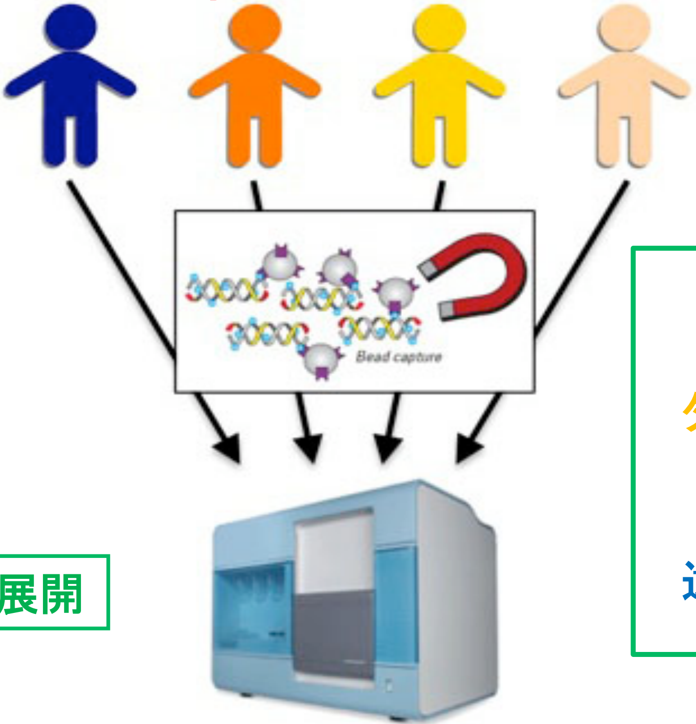
RNA

DNA

Next Gen Sequencing Analysis

Data compilation

次世代シークエンサーを用いた解析基盤の確立



臨床検体・実験試料

ゲノム解析

RNA解析

手法の開発と応用

ゲノム構造の解析

タンパク質結合部位の解析

転写開始点の解析

遺伝子発現量の解析

様々な生物種、実験系への展開

次世代のゲノム解析へ

ACKNOWLEDGEMENTS

イルミナの運用とデータ基礎解析：

菅野研（東大）

DBTSSの作成と解析：

中井研（東大）

マウス性細胞のメチル化（全ゲノムバイサルファイト）：

河野研（農大）

質量分析計を用いた解析：

夏目研（JBIC）

イルミナ解析技術の開発：

金井（広大）

ガン組織リシークエンス：

江角研（がんセンター東病院）

免疫担当細胞のChIP Seq：

田中研（千葉大）

選択的スプライスの解析：

今西研（JBIC）