

Genomon2 ハンズオンセッション 実践編

東京大学医科学研究所ヒトゲノム解析センター

千葉健一
岡田愛
白石友一

Genomon2 ハンズオンセッション

このハンズオンセッションは、Linuxのコマンドが使えるユーザ様を対象に、Genomon2の解析、結果の確認、そしてパラメータのチューニングを行えるようになることを目的としています。

本発表の流れ

- GenomonDNA解析の実行方法について説明
- DNA解析を実行してみましよう！（実習）
- DNA解析結果を確認しましょう。
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
- RNA解析を実行してみましよう！（実習）
- RNA解析結果を確認しましょう。
- paplotの説明（RNA）

本発表の流れ

- **GenomonDNA解析の実行方法について説明**
 - Genomon実行コマンドについて説明
 - サンプル設定ファイルの解説
 - パイプライン設定ファイルの解説
- DNA解析を実行してみましよう！（実習）
- DNA解析結果を確認しましょう。
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
- RNA解析を実行してみましよう！（実習）
- RNA解析結果を確認しましょう。
- paplotの説明（RNA）

Genomonの実行 HGCスパコン用

```
$ bash /share/pub/genomon/genomon_pipeline-  
2.6.3/genomon_script/genomon_pipeline_HGC.sh (解析タイプ) (サンプル設定  
ファイル) (出カルートディレクトリ) (パイプライン設定ファイル)
```

引数1: 解析タイプ

DNA解析をする場合は, dna を, RNA解析をする場合は, rna を指定します.

引数2: サンプル設定ファイル

解析対象のファイル (FASTQやBAM) のパスなどを記載したファイルを作成し, そのファイル名を指定します.

今日の実習ではこちらで用意したサンプル
5929の解析をしていただきます

引数3: 出カルートディレクトリ

出カルートディレクトリを指定します. 結果がこのディレクトリ以下に出力されます.

引数4: パイプライン設定ファイル

DNA解析用のパイプライン設定ファイルを指定します.

HGCスパコンに既に用意されているパイプライン
設定ファイルをコピーして使用します

5929のサンプル設定ファイルの解説

[fastq]

5929_tumor1, /path/to/tumor/sequence1.fastq, /path/to/tumor/sequence2.fastq

(tumor2~4 省略)

5929_control1, /path/to/control/sequence1.fastq, /path/to/control/sequence2.fastq

(control2~4 省略)

5929_pool1, /path/to/pool/sequence1.fastq, /path/to/pool/sequence2.fastq

(control2~10 省略)

[sv_detection]

5929_tumor1, 5929_control1, panel1

[mutation_call]

5929_tumor1, 5929_control1, panel1

[qc]

5929_tumor1

5929_control1

[controlpanel]

panel1, 5929_pool1, 5929_pool2, 5929_pool3, 5929_pool4, 5929_pool5, 5929_pool6, 5929_pool7, 5929_pool8, 5929_pool9, 5929_pool10

[fastq]

(サンプル名), (FASTQ1ファイルパス), (FASTQ2ファイルパス)

サンプル名ですが、アルファベット、数値、ハイフン、アンダーバーは使用できます。あまり特殊な文字を使用しないでください。

入力情報タグで指定したサンプル名を使用して、解析情報タグを作成します。

[sv_detection]

(Tumorサンプル名), (Normalサンプル名), (コントロールパネル名)

[mutation_call]

(Tumorサンプル名), (Normalサンプル名), (コントロールパネル名)

[qc]

(サンプル名)

[controlpanel]

(コントロールパネル名), (Normalサンプル名1), (Normalサンプル名2) ... (Normalサンプル名N)

*controlpanelがない場合は、Noneを指定します。

サンプル設定ファイルの項目の説明

• 入力情報タグ

入力ファイルの情報を記載します

[fastq] FASTQファイルを入力として解析します

[bam_tofastq] BAMを入力してFASTQファイルに戻してから解析します

[bam_import] 入力したBAMを再マッピングせずに解析を実行します (Genomon2でアライメントしたBAMファイルが対象です)

• 解析情報タグ

解析するための情報を記載します

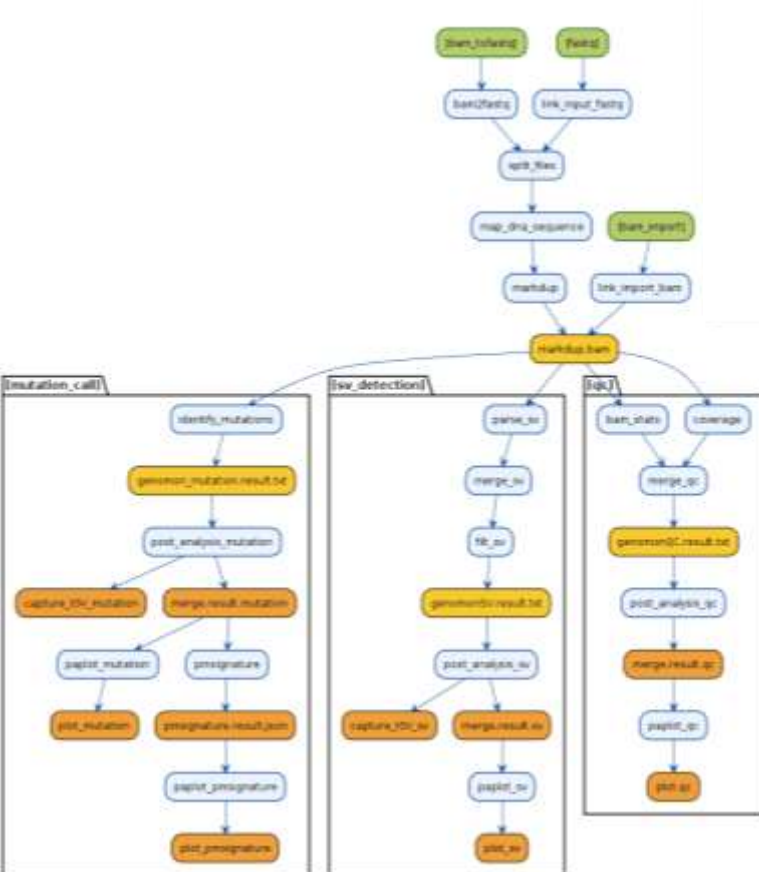
[mutation_call] 変異コールが実行されます

[sv_detection] SV検出が実行されます

[qc] BAMのQuality Controlを出力します

[controlpanel] コントロールパネルを指定します

Genomon Pipeline Schemes DNA



パイプライン設定ファイル DNAの解説 (advanced)

[REFERENCE]

prepared reference fasta file

ref_fasta = /xxx/database/GRCh37/GRCh37.fa

interval_list = /xxx/database/GRCh37/GRCh37_noScaffold_noDecoy.interval_list

...(省略) ...

[REFERENCE]には

使用するリファレンスゲノムやデータベースが記載されています

[SOFTWARE]

prepared tools

blat = /xxx/install/blat_x86_64/blat

bwa = /xxx/install/bwa-0.7.8/bwa

...(省略) ...

[SOFTWARE]には

使用するソフトウェアのパスが記載されています

#####

parameters for split fastq

[split_fastq]

qsub_option = -l s_vmem=1G,mem_req=1G

split_fastq_line_number = 40000000

fastq_filter = False

[split_fastq],[bwa_mem]などでは、

各ステージで使用するパラメータやメモリ量などを調整できます

#####

parameters for bwa_mem

[bwa_mem]

qsub_option = -l s_vmem=10.6G,mem_req=10.6G

bwa_params = -T 0

HGCに用意されているパイプライン設定ファイルは、基本的な解析に適したパラメータの値となっておりますので、まずは、値を変更しないで動かして、出力結果をみて調整しましょう

パラメータの調整については、本日の講習に含まれます

...(以下略) ...

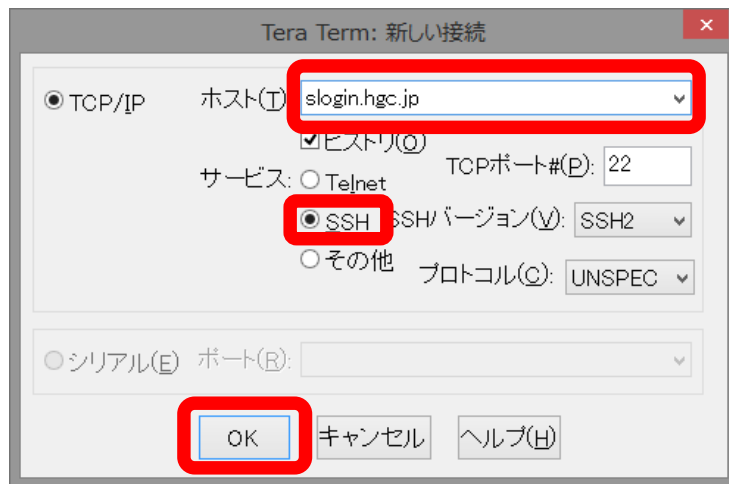
本発表の流れ

- GenomonDNA解析の実行方法について説明
- **DNA解析を実行してみましよう！**
 - スパコンへのログインとコマンドの説明
 - **GenomonでDNA解析（実習）**
- DNA解析結果を確認しましょう。
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
- RNA解析を実行してみましよう！（実習）
- RNA解析結果を確認しましょう。
- paplotの説明（RNA）

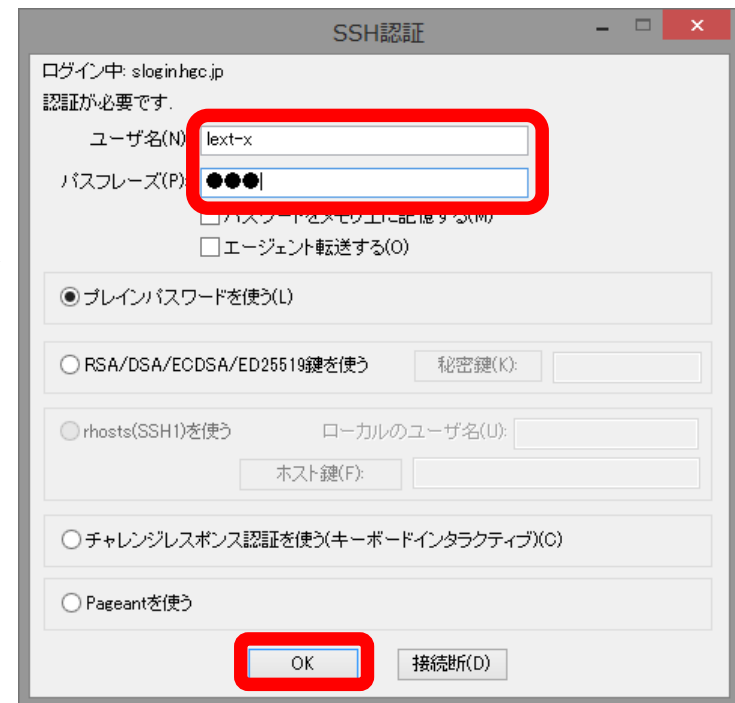
Windows からのログイン Tera Term を使用する場合 (パスワード認証)

1. slogin.hgc.jp にログインしてください

Windows (Tera Term 使用) の場合

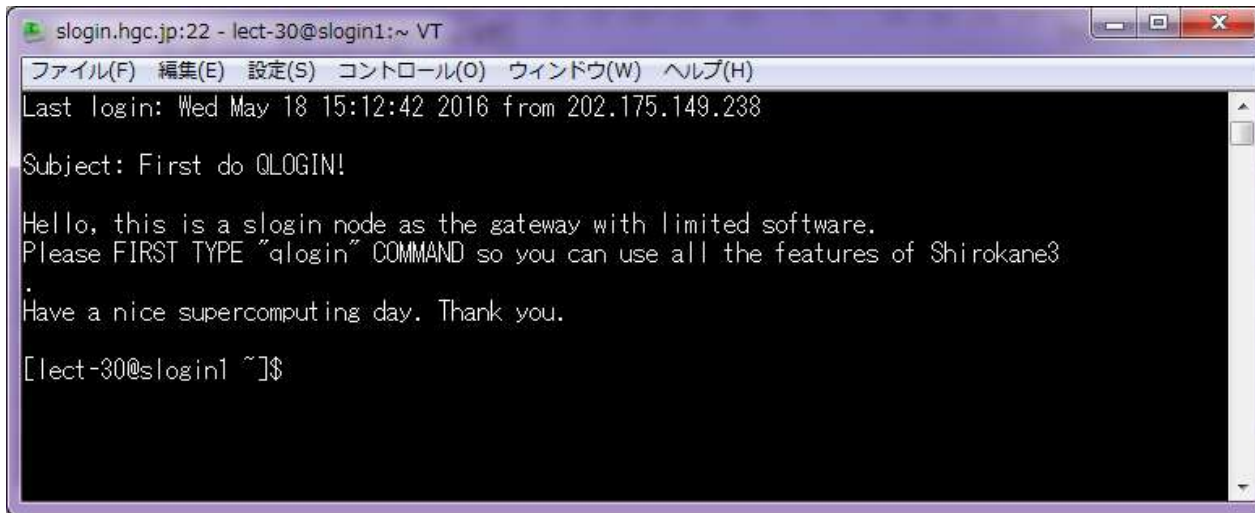


- ・ ホスト : **slogin.hgc.jp** を入力
- ・ サービス : **SSH** を選択
- OK ボタンをクリック



- ・ ユーザ名 : **lect-x** を入力
- ・ パスフレーズ : **現在のパスワード** を入力
- OK ボタンをクリック

ログイン時の画面表示



The screenshot shows a terminal window titled "slogin.hgc.jp:22 - lect-30@slogin1:~ VT". The menu bar includes "ファイル(F)", "編集(E)", "設定(S)", "コントロール(O)", "ウィンドウ(W)", and "ヘルプ(H)". The terminal text reads: "Last login: Wed May 18 15:12:42 2016 from 202.175.149.238", "Subject: First do QLOGIN!", "Hello, this is a slogin node as the gateway with limited software. Please FIRST TYPE 'qlogin' COMMAND so you can use all the features of Shirokane3.", "Have a nice supercomputing day. Thank you.", and the prompt "[lect-30@slogin1 ~]\$".

[lect-x@slogin1~]\$ █ ← カーソル

ユーザ名 ホスト名 カレントディレクトリ
(チルダ(～)は、ホームディレクトリを
示します)

ユーザがコマンドを実行できる状態

※ただし、システムやシェルによって表示内容は異なります。

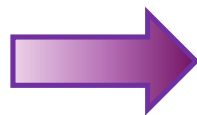
ログインノード

インターネットから、スパコンに接続する際は、まずログインノードに接続します。外部からの接続を限定することで、セキュリティを保っています。

インターネット



ログインノード



slogin.hgc.jp



% qlogin

計算ノード(スパコン)



ログインノードは、ログインおよびインターネットからスパコンへのデータ転送に特化したノードです。ジョブ実行や、スパコンからインターネットへのデータ転送は、計算ノードに qlogin してから行ってください。

計算ノードへのログイン

qlogin します。(qlogin と入力)

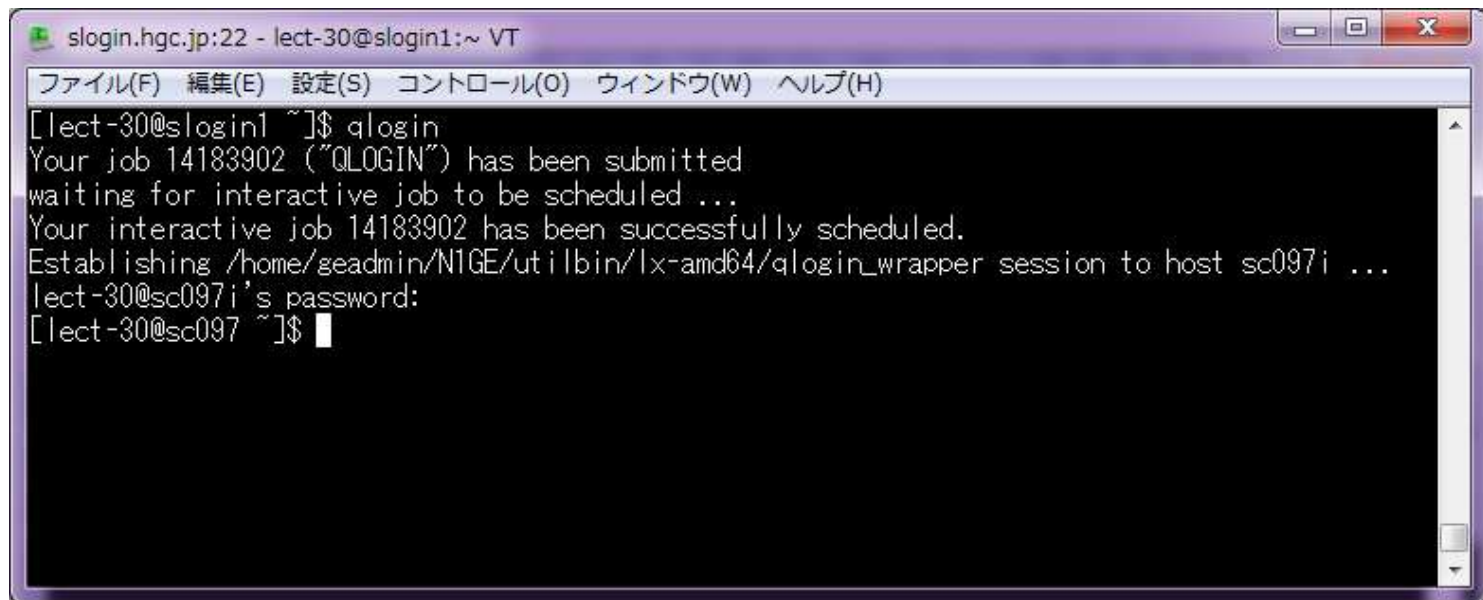
```
[lect-x@slogin1 ~]$ qlogin
```

```
lect-x@sc097i's password: ←現在のパスワードを入力します
```

qlogin すると自動的にいずれかの計算ノードに割当てられます。鍵交換認証をしている場合は自動でログインします。

(lect userはパスワード認証のため、パスワードを入力してログインします。)

qlogin時の画面表示



```
slogin.hgc.jp:22 - lect-30@slogin1:~ VT
ファイル(F) 編集(E) 設定(S) コントロール(O) ウィンドウ(W) ヘルプ(H)
[lect-30@slogin1 ~]$ qlogin
Your job 14183902 ("QLOGIN") has been submitted
waiting for interactive job to be scheduled ...
Your interactive job 14183902 has been successfully scheduled.
Establishing /home/geadmin/NTGE/utilbin/lx-amd64/qlogin_wrapper session to host sc097i ...
lect-30@sc097i's password:
[lect-30@sc097 ~]$
```

[lect-x@sc097~]\$ █ ← カーソル

ユーザ名 ホスト名 カレントディレクトリ
(チルダ(~)は、ホームディレクトリを示します)

計算ノードにログインしている状態です！

ジョブの実行状況を確認する

投入したジョブの状況は `qstat` コマンドで確認します。

① 待ち行列にジョブが入る。

```
[lect-x@sc097 ~]$ qstat
```

| job-ID | prior | name | use | state | submit/start at | queue | slots | ja-task-ID |
|----------|---------|------------|--------|-------|---------------------|-------|-------|------------|
| 14238841 | 0.00004 | qsub_genom | lect-x | qw | 05/19/2016 16:40:02 | | 1 | |

\$

state に「qw」が表示される。

② ジョブが実行ホストに転送される（短い）

```
[lect-x@sc097 ~]$ qstat
```

| job-ID | prior | name | use | state | submit/start at | queue | slots | ja-task-ID |
|----------|---------|------------|--------|-------|---------------------|----------------|-------|------------|
| 14238841 | 0.00004 | qsub_genom | lect-x | t | 05/19/2016 16:42:41 | ljobs.q@sc432i | 1 | |

\$

state に「t」が表示される。

③ ジョブが実行される

```
[lect-x@sc097 ~]$ qstat
```

| job-ID | prior | name | use | state | submit/start at | queue | slots | ja-task-ID |
|----------|---------|------------|--------|-------|---------------------|----------------|-------|------------|
| 14238841 | 0.00004 | qsub_genom | lect-x | r | 05/19/2016 16:42:41 | ljobs.q@sc432i | 1 | |

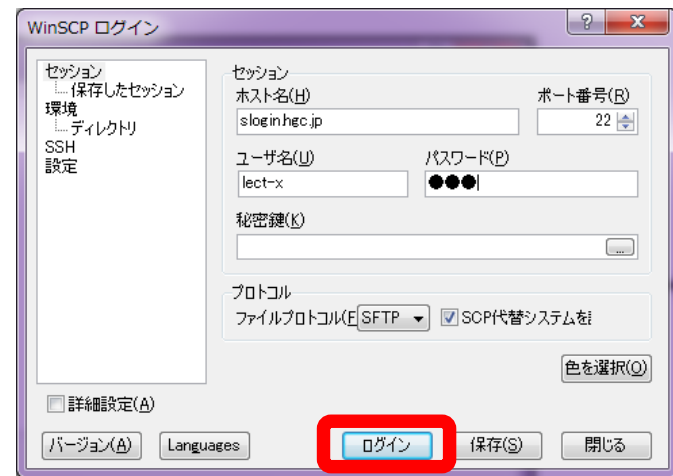
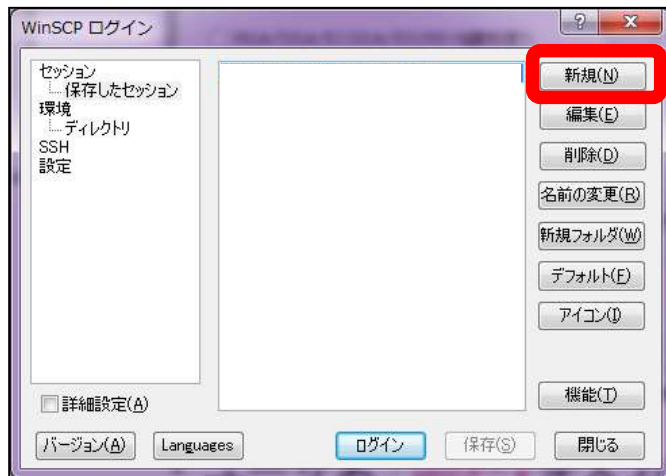
\$

state に「r」が表示される。

Windows からのファイルダウンロード WinSCP を使用する場合 (パスワード認証)

1. slogin.hgc.jp にログインしてください。

Windows (WinSCP 使用) の場合



→ 新規 ボタンをクリック

下記の警告が出たら はい(Y)ボタンをクリック



- ・ ホスト : **slogin.hgc.jp** を入力
 - ・ ユーザ名 : **lect-x** を入力
 - ・ パスワード : **現在のパスワード** を入力
 - ・ サービス : **SFTP** を選択
- ログインボタンをクリック

実習！

#1 Tera TermでHGCログインノードにログインする (windowsのデスクトップにTeraTermがあります)

#2 HGCスパコンにログインする

\$ **qlogin**

#3 作業用ディレクトリの作成

\$ **mkdir Genomon2_6_3/**

\$ **mkdir Genomon2_6_3/config/**

#4 こちらで用意したサンプル設定ファイルを、作成したディレクトリにコピーする

\$ **cp /share/pub/genomon/genomon_pipeline-2.6.3/sample_sheet/test_dna/5929_sample.csv Genomon2_6_3/config**

#5 こちらで用意したパイプライン設定ファイル(exome用)を、作成したディレクトリにコピーする

\$ **cp /share/pub/genomon/genomon_pipeline-2.6.3/genomon_conf/dna_exome_genomon.cfg Genomon2_6_3/config**

#6 Genomon2の実行 (以下、3行に分かれていますが、一行のコマンドです。一度に実行してください)

\$ **bash /share/pub/genomon/genomon_pipeline-2.6.3/genomon_script/genomon_pipeline_HGC.sh dna**

Genomon2_6_3/config/5929_sample.csv Genomon2_6_3/test5929_exome

Genomon2_6_3/config/dna_exome_genomon.cfg

#7 Jobが実際に動いているか確認する

\$ **qstat** (watch qstat コマンドでモニタリングできます。完了したらCtrl+cでキャンセルしましょう)

Jobが完了したら次のページへ

#8 WinSCPを使用して以下のファイルを任意のディレクトリにダウンロードしましょう

QC結果

[Genomon2_6_3/test5929_exome/post_analysis/5929_sample/merge_qc.txt](#)

変異コール結果

[Genomon2_6_3/test5929_exome/post_analysis/5929_sample/merge_mutation.txt](#)

[Genomon2_6_3/test5929_exome/post_analysis/5929_sample/merge_mutation_filt.txt](#)

SV結果

[Genomon2_6_3/test5929_exome/post_analysis/5929_sample/merge_sv.txt](#)

[Genomon2_6_3/test5929_exome/post_analysis/5929_sample/merge_sv_filt.txt](#)

paplot結果(ディレクトリごとダウンロード)

[Genomon2_6_3/test5929_exome/paplot](#)

#9 QC結果、変異コール結果、SV結果のテキストファイル(タブ区切りになってます)を、OpenOfficeやExcelなどのソフトで表示する

#10 paplot/5929_sample/index.htmlをダブルクリックして表示する

#11

#ANNOVARを使用した変異コール結果を用意したので表示してみましよう

ダウンロードしたディレクトリの中にあります

[20171030/merge_mutation_hands-on171030.txt](#)

本発表の流れ

- GenomonDNA解析の実行方法について説明
- DNA解析を実行してみましょう！
- **DNA解析結果を確認しましょう。**
 - Bam QCの確認
 - 変異コール結果の確認
 - SV検出結果の確認
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
- RNA解析を実行してみましょう！（実習）
- RNA解析結果を確認しましょう。
- paplotの説明（RNA）

baitファイルについて

BaitファイルとはExomeやTargetシーケンスした際の、DNAをキャプチャした領域が記載されているファイルのことを指します。

BAMのクオリティーをチェックするために、baitファイルをパイプライン設定ファイルに指定する必要があります。

average_depth

キャプチャした領域に対して、average_depthがどのくらい算出しています

10x, 20x, 30x_ratio

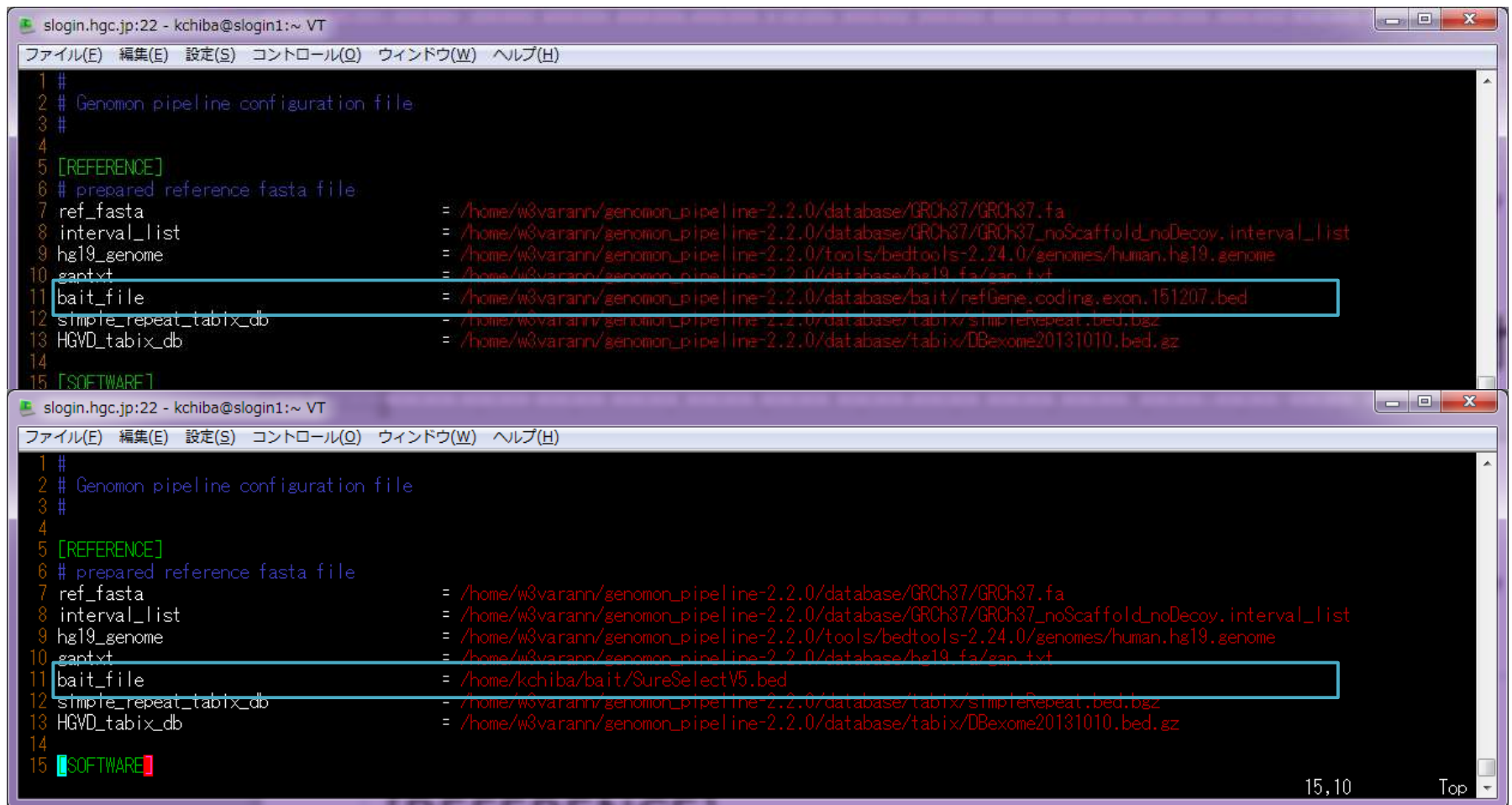
キャプチャした領域に対して、どの程度のdepthがカバーされているか算出しています

パイプライン設定ファイルの”bait_file”を、ユーザ様がご使用したSureSelect,Truseqなどのキャプチャ領域が記載されているファイルに変更してください。

baitファイルの変更方法

パイプライン設定ファイルの
[REFERENCE]

bait_file = に指定しているファイルを変更します.



The image shows two screenshots of a terminal window with a purple title bar, demonstrating how to change the `bait_file` in a Genomon pipeline configuration file. The terminal window has a menu bar with options: ファイル(F), 編集(E), 設定(S), コントロール(O), ウィンドウ(W), ヘルプ(H).

Top Screenshot: The configuration file is open, showing the `[REFERENCE]` section. The `bait_file` is currently set to `/home/w3varann/genomon_pipeline-2.2.0/database/bait/refGene.coding.exon.151207.bed`. This line is highlighted with a blue box.

```
1 #
2 # Genomon pipeline configuration file
3 #
4
5 [REFERENCE]
6 # prepared reference fasta file
7 ref_fasta = /home/w3varann/genomon_pipeline-2.2.0/database/GRCh37/GRCh37.fa
8 interval_list = /home/w3varann/genomon_pipeline-2.2.0/database/GRCh37/GRCh37_noScaffold_noDecoy.interval_list
9 hg19_genome = /home/w3varann/genomon_pipeline-2.2.0/tools/bedtools-2.24.0/genomes/human.hg19.genome
10 gartxt = /home/w3varann/genomon_pipeline-2.2.0/database/hg19.fa/gartxt
11 bait_file = /home/w3varann/genomon_pipeline-2.2.0/database/bait/refGene.coding.exon.151207.bed
12 simple_repeat_tabix_db = /home/w3varann/genomon_pipeline-2.2.0/database/tabix/simpleRepeat.bed.bgz
13 HGVD_tabix_db = /home/w3varann/genomon_pipeline-2.2.0/database/tabix/DBexome20131010.bed.gz
14
15 [SOFTWARE]
```

Bottom Screenshot: The same configuration file is shown, but the `bait_file` has been changed to `/home/kchiba/bait/SureSelectV5.bed`. This line is also highlighted with a blue box.

```
1 #
2 # Genomon pipeline configuration file
3 #
4
5 [REFERENCE]
6 # prepared reference fasta file
7 ref_fasta = /home/w3varann/genomon_pipeline-2.2.0/database/GRCh37/GRCh37.fa
8 interval_list = /home/w3varann/genomon_pipeline-2.2.0/database/GRCh37/GRCh37_noScaffold_noDecoy.interval_list
9 hg19_genome = /home/w3varann/genomon_pipeline-2.2.0/tools/bedtools-2.24.0/genomes/human.hg19.genome
10 gartxt = /home/w3varann/genomon_pipeline-2.2.0/database/hg19.fa/gartxt
11 bait_file = /home/kchiba/bait/SureSelectV5.bed
12 simple_repeat_tabix_db = /home/w3varann/genomon_pipeline-2.2.0/database/tabix/simpleRepeat.bed.bgz
13 HGVD_tabix_db = /home/w3varann/genomon_pipeline-2.2.0/database/tabix/DBexome20131010.bed.gz
14
15 [SOFTWARE]
```

In the bottom screenshot, the status bar at the bottom right shows "15,10" and "Top".

BAM QCの結果を確認する

シーケンスの実験がうまくいっているのかの確認を行います。

確認する項目：

`#_total_reads`, `#_mapped_reads`, `#_duplicate_reads`

`average_depth`:

WGS 30x～50x , WES 100x～, targeted seq 500x～

10x, 20x, 30x_ratio: WGS,WES

50x, 100x_ratio: targeted seq

変異コールの結果を確認する

- ・変異

`merge_mutation.filt.txt`

- 変異コール結果. P値などで適切なフィルタ済み

`merge_mutation.txt`

- 変異コール結果. フィルタなしデータ
(こちらはadvanced user向け)

フィルタ済みのファイルもありますが、フィルタが解析サンプルに対して適切かある程度評価する必要があります。

ANNOVARのインストールについて

今回の実習ではANNOVARによるアノテーションを実施しませんでした。

ANNOVARを使用しないと、変異のポジションの遺伝子名や、そこがCoding領域かどうか、SNPであるか、などという情報が出力されません。

ANNOVARは各ユーザ様でインストールする必要があります。

ANNOVARのインストールについてはGenomonのマニュアルに詳細が記載されていますので、Genomonを実際に使われる際には是非インストールしてください。

http://genomon.readthedocs.io/ja/latest/dna_quick_start.html

このあとの変異コールの説明では、ANNOVARでアノテーションしないと出てこない項目についても、説明いたします。

Coding領域をチェックする

まずは、coding領域の変異をみる。

coding領域はdriver変異や症例あたりの変異数が知られているので、解析がうまくいっているのか解釈しやすい。

症例ごとの変異の数や変異アレル頻度、変異のタイプなどをチェック。

→あるはずの変異が見つからない(例: 膵臓がんでのKRAS etc.)、

変異数が少なすぎる or 多すぎる場合はfilterの再検討やサンプルの質に問題がないかの確認を行う。

result.txt から result.filt.txt は どのようなフィルタを行っているか

BAM



①変異候補をゆるめの閾値で検出する

`${sample}.genomon_mutation.result.txt`



②おすすめフィルタで変異候補を絞る

`${sample}. genomon_mutation.result.filt.txt`

・あるはずの変異がresult.filt.txtにない！

→まずは、result.txtに変異候補がないか確認しましょう。

・result.txtにあった

→result.filt.txtに出力されない原因を確認しましょう（Normalに変異がはいっているため検出されないとか）

パイプライン設定ファイルを変更して ②のおすすめフィルタを変更する

おすすめフィルタのコマンド

/path/to/mutil filter -i \$input.txt -o \$output.txt ¥

{pair_params}

[fisher_mutation_call]

pair_params = --fish_pval 1.0 --realign_pval 1.0 --eb_pval 4.0 --
tcount 4 --ncount 2

--fish_pval: カラム"P-value(fisher)" >= 1.0

--realign_pval: カラム"P-value(fisher)_realignment" >= 1.0

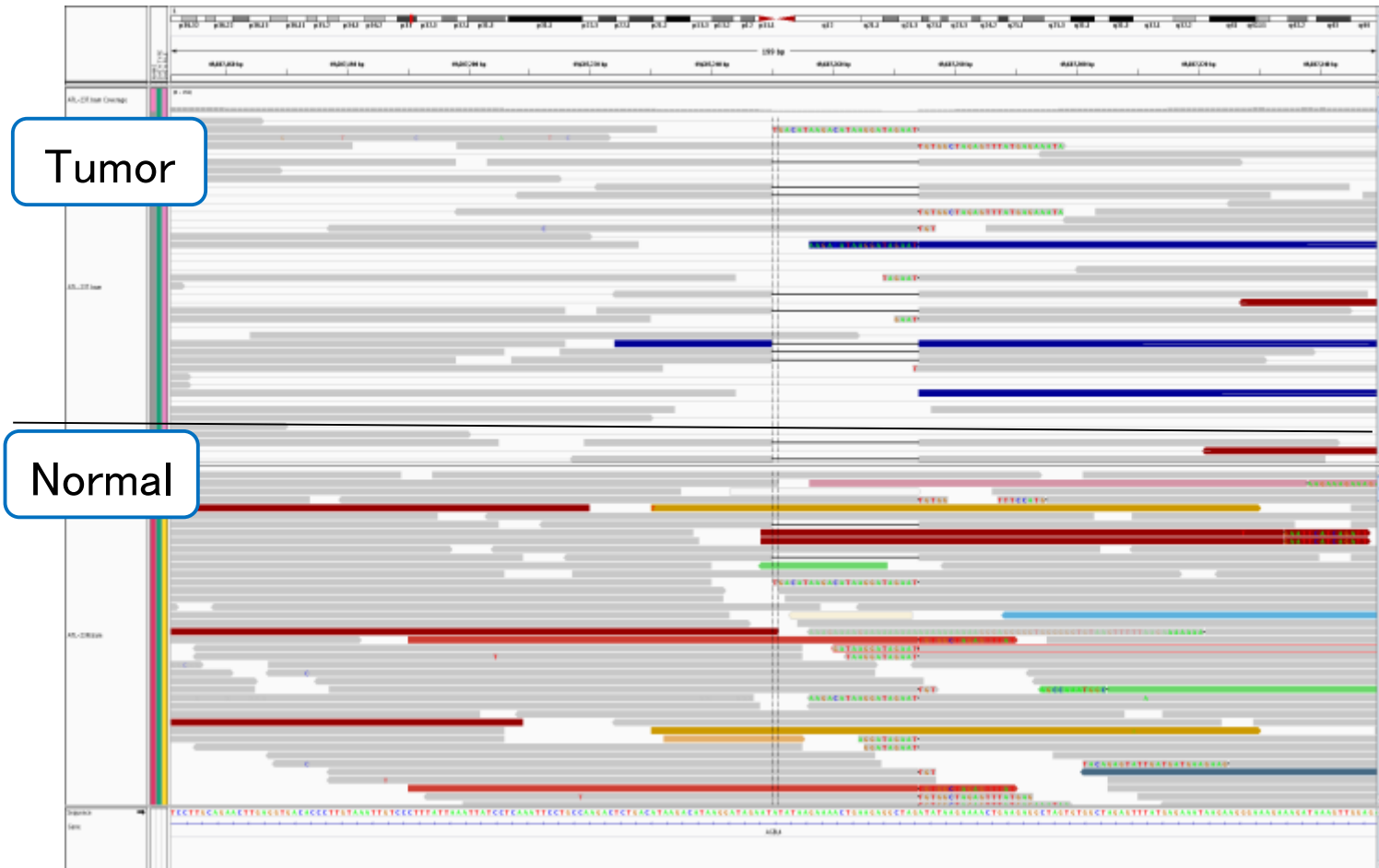
--eb_pval: カラム"P-value(EBCall)" >= 4.0

--tcount: カラム"AltNum_tumor" >= 4

--ncount: カラム"AltNum_normal" <= 2

Excelのフィルタ機能と同じイメージです。

local realignmentについて

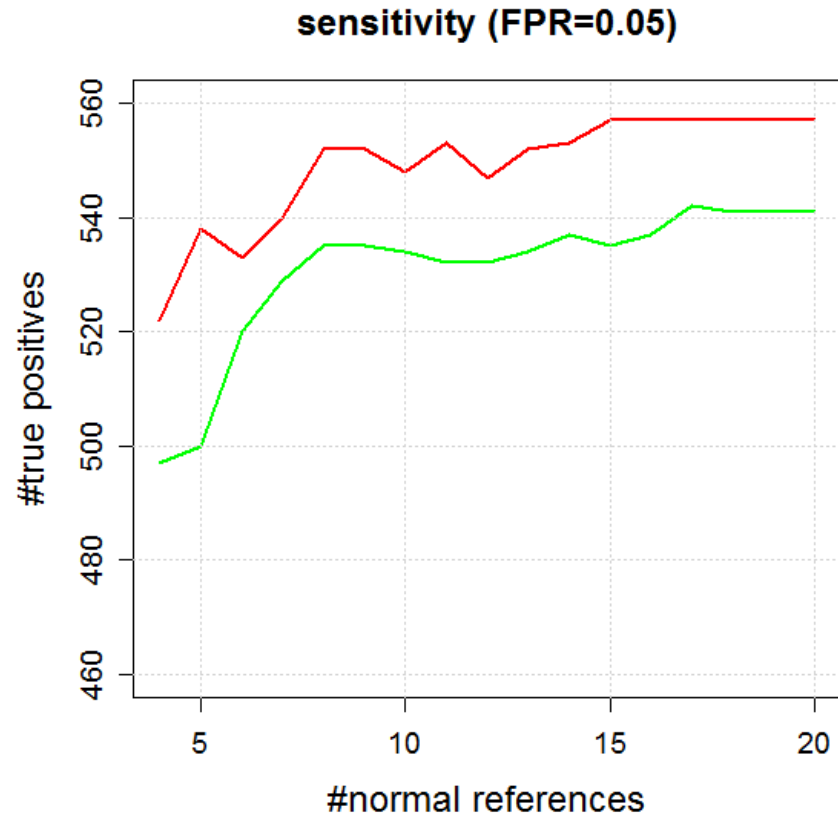


variantPairNum_tumor, variantPairNum_normalという項目が、local realignmentして Variantリードの数を再計算した値になります。それらの値でFisher検定した結果がP-value(fisher_realignment)になります。

EBCallのcontrol panel数

- Control Panelのサンプル数
 - 推奨は20
 - $P\text{-value(EBCall)} = 4.0$
 - ・10サンプルくらいでもFisher検定のみよりはずっと良い
 - $P\text{-value(EBCall)} = 3.0$

あとは解析対象疾患に合うように 解析する人が最適な基準を設定しましょう！



赤: 同じコホートの検体
緑: 別のコホートの検体

Shiraishi et al. NAR, 2013

その他のフィルタ

filtering SNP filter (dbSNP 、1000 genome project 、ESP 、HGVD ...)

strand ratio 0 or 1

SNP filter

SNP databaseで変異候補をフィルタすることは有用です.

しかしdnSNP131などのデータベースにあっても低頻度なもののや、特定の遺伝子やCOSMICにあるものなどはdriver変異である場合があるので、それらを確認してフィルタしましょう.

- dbSNP131

JAK2 V617F

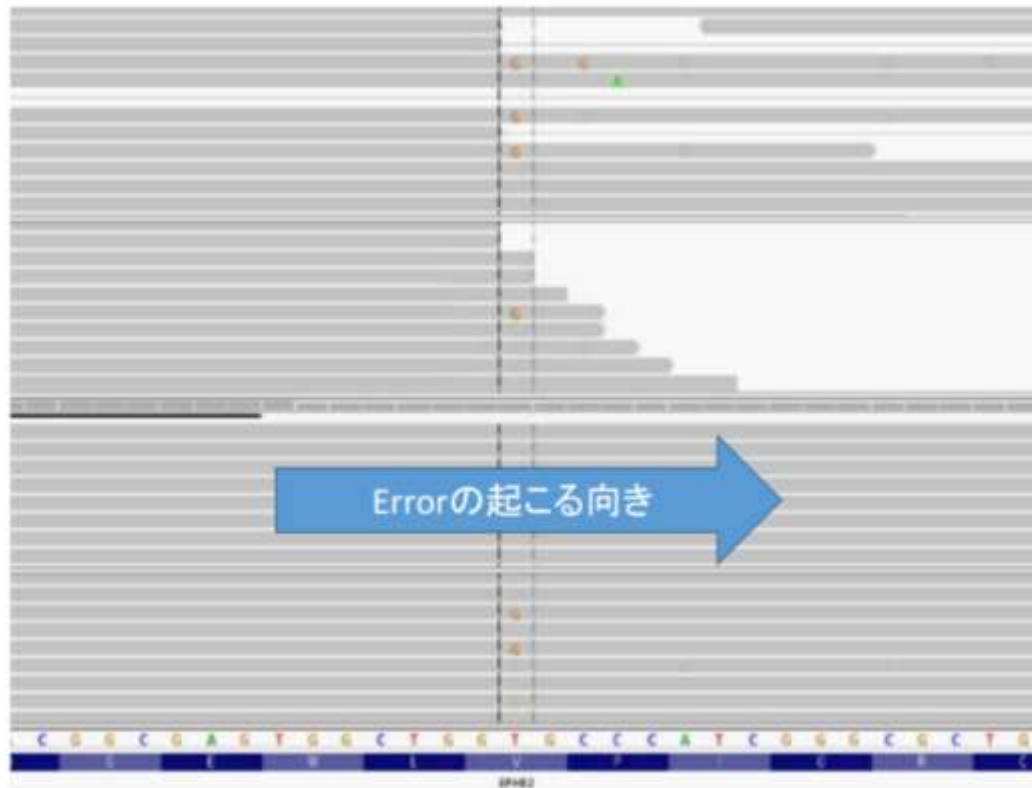
FANCA frameshift mutation

(dbSNP135はもっと多い)

- ESP5400

DNMT3A hot spot mutation

Mutations read in only one direction



連続した塩基配列の後は読み間違いが起こりやすい(特定の方向でのみエラー)

→strand_ratioが0あるいは1になる.

ただし、Exomeなどで遺伝子をキャプチャしている場合、キャプチャした領域の端の部分はstrandがかたよりやすい。strand_ratioが0あるいは1だからといって、確実にエラーだと判断できるわけではないので注意が必要.

SV検出の結果を確認する

- ・SV検出

`merge_sv.filt.txt`

– SV検出結果. 適切なフィルタ済み

`merge_sv.txt`

–SV検出結果. ゆるめのフィルタデータ

(こちらはadvanced user向け)

フィルタ済みのファイルもありますが、フィルタが解析サンプルに対して 適切かある程度評価する必要があります。

result.txt から result.filt.txt は どのようなフィルタを行っているか

BAM



①SV候補をゆるめの閾値で検出する

`${sample}. genomonSV.result.txt`



②おすすめフィルタでSV候補を絞る

`${sample}. genomonSV.result.filt.txt`

・あるはずの変異がresult.filt.txtにない！

→まずは、result.txtに変異候補がないか確認しましょう。

・result.txtにあった

→result.filt.txtに出力されない原因を確認しましょう（Normalに変異がはいっているため検出されないとか）

パイプライン設定ファイルを変更して ②のおすすめフィルタを変更する

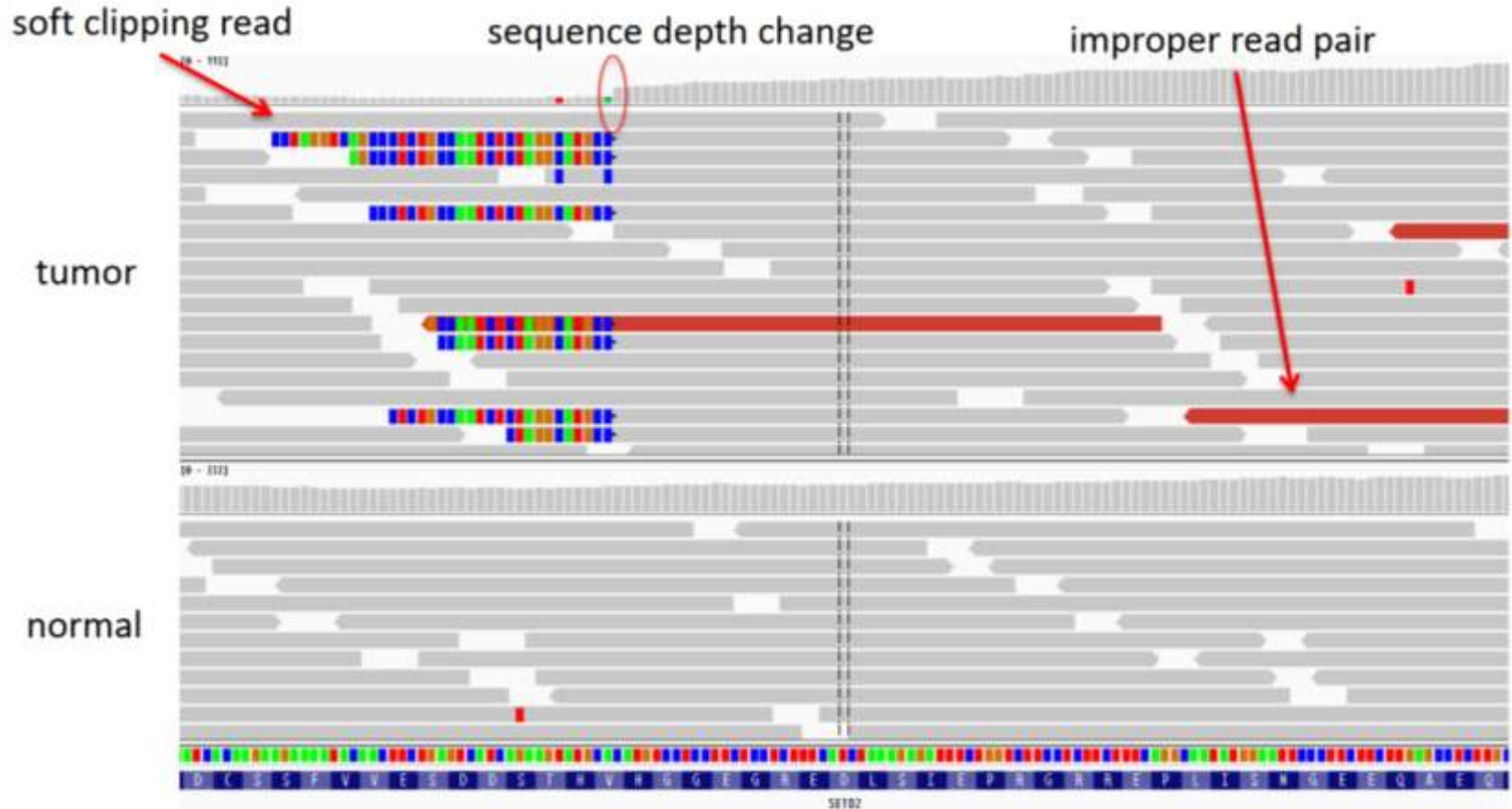
```
# おすすめフィルタのコマンド  
/path/to/sv_utils filter ¥  
input.txt output.txt $sv_utils_annotation_dir ¥  
{sv_utils_param}
```

[sv_filt]

```
sv_utils_params = --min_tumor_allele_freq 0.07 --  
max_control_variant_read_pair 1 --control_depth_thres 10 --  
inversion_size_thres 1000 --remove_simple_repeat  
  
--min_tumor_allele_freq 0.07  
--max_control_variant_read_pair 1  
--control_depth_thres 10  
--inversion_size_thres 1000
```

Excelのフィルタ機能と同じイメージです。

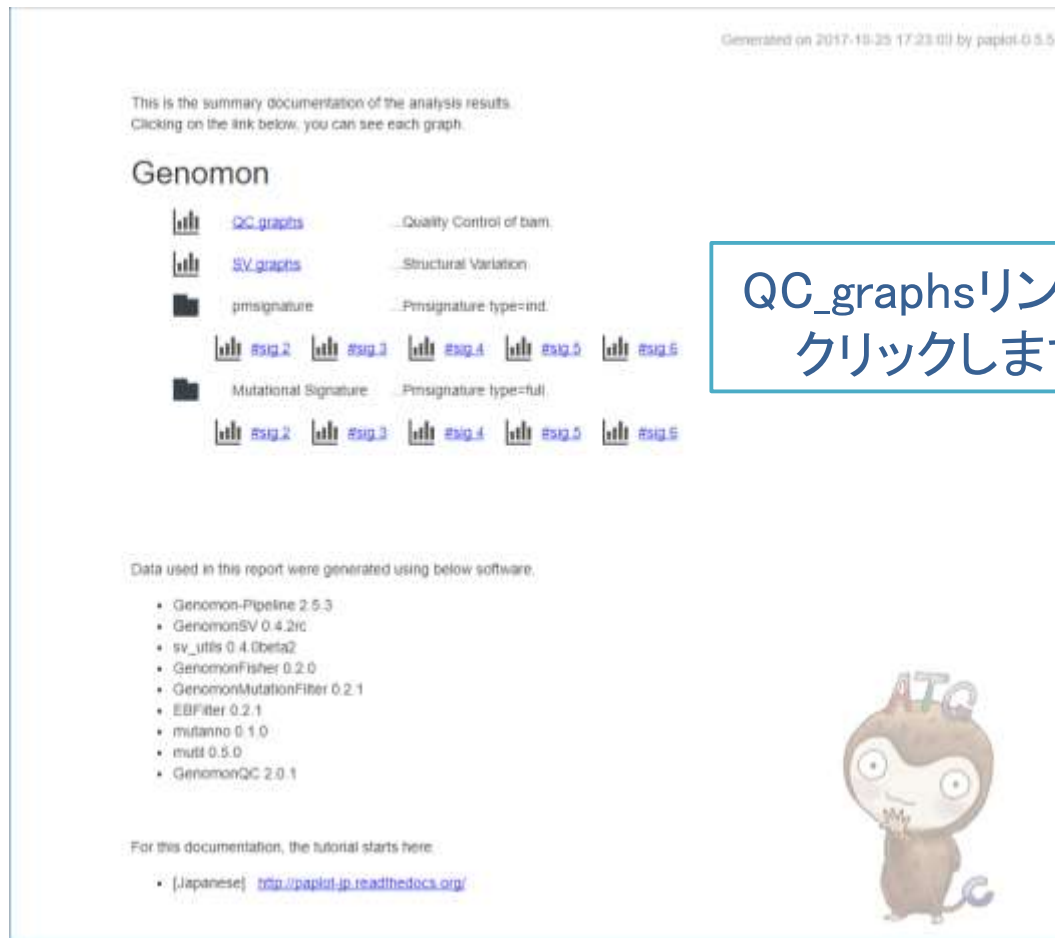
SVが本物かチェックする



- この場合Soft clipping read, improper read pairは、chr2:286238293付近に張り付く
- Sequence depthが変化するとは限らない
- controlにはこれらの変化がないということも非常に重要

paplotでQCを確認する

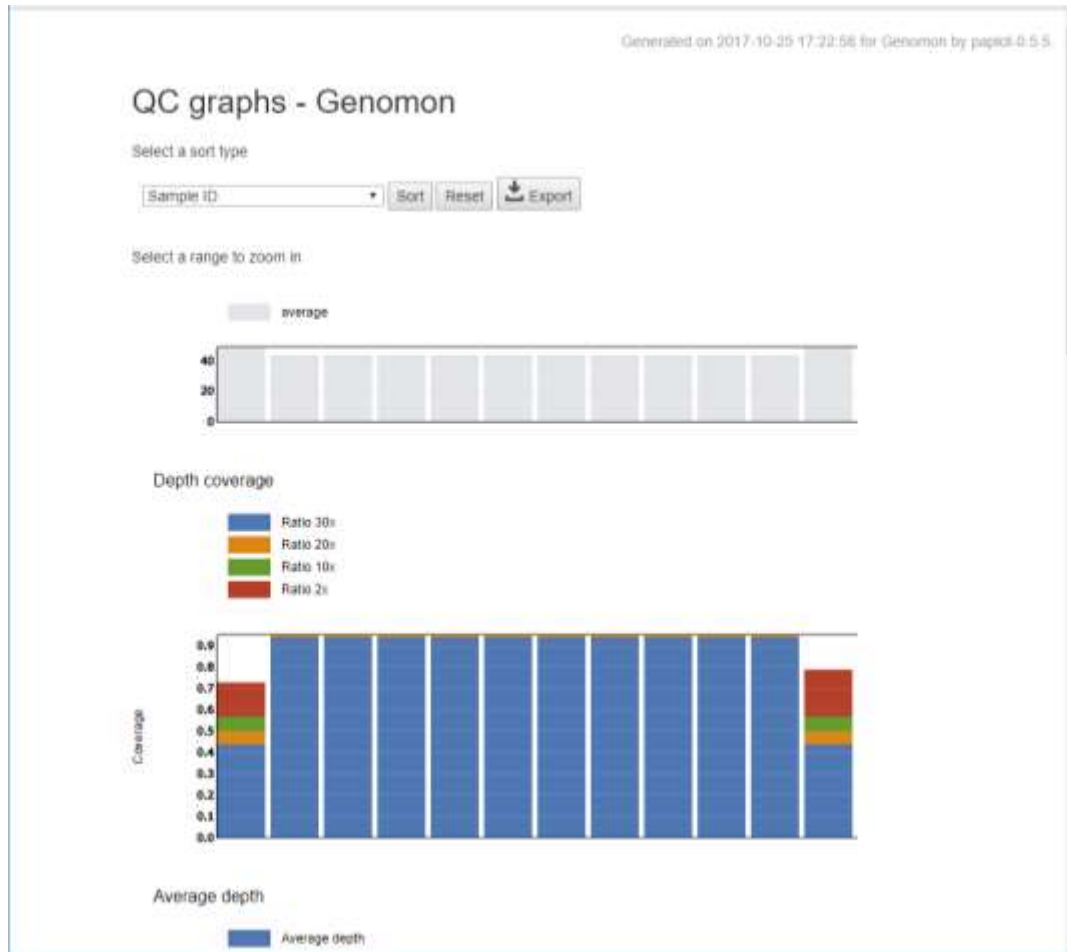
ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリックしてください



QC_graphsリンクを
クリックします

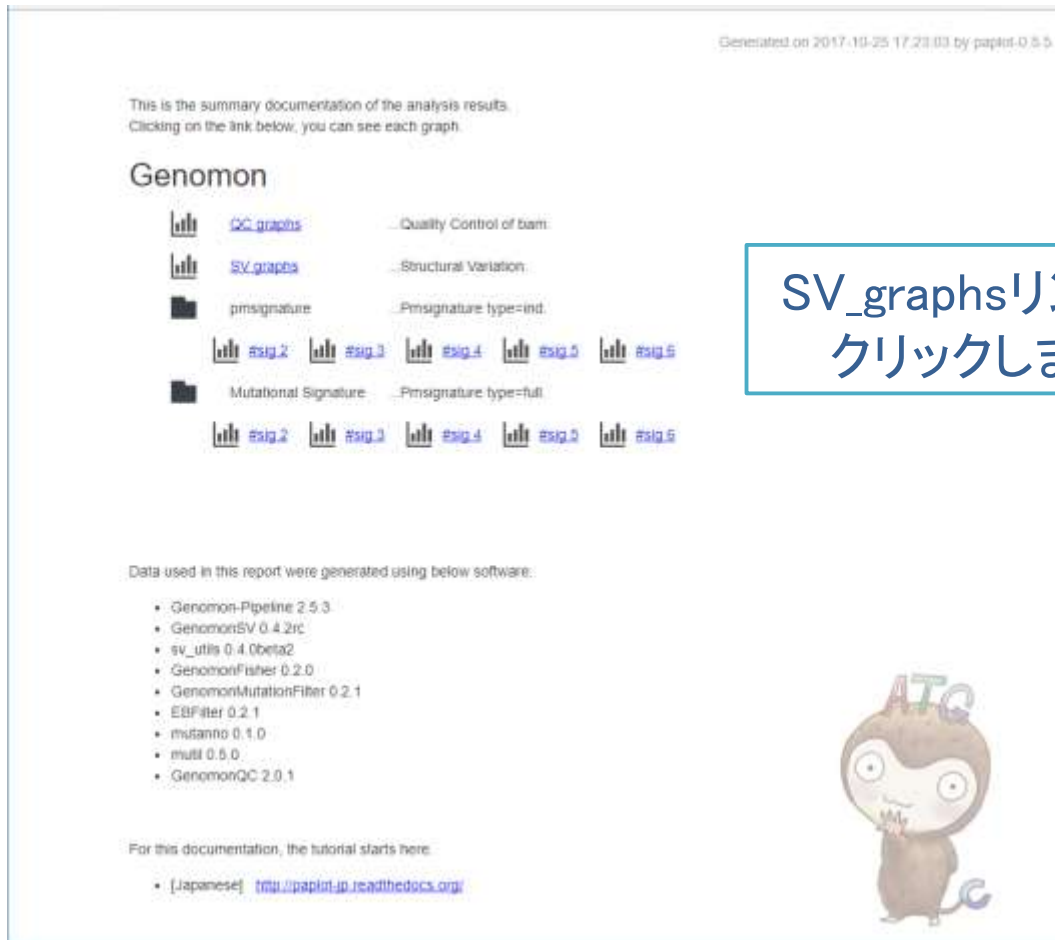
paplotでQCを確認する

Sample 5929の結果が表示されます.



paplotでSV結果を確認する

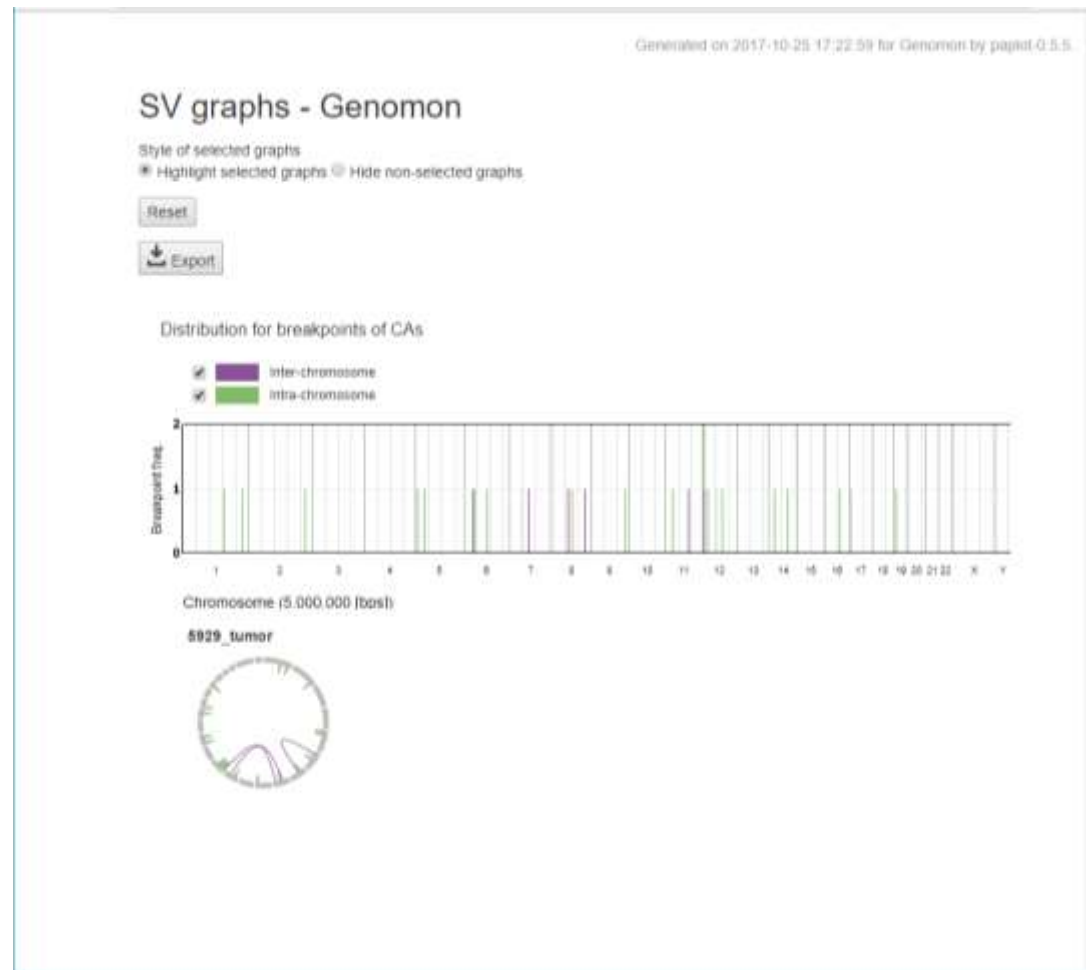
ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリックしてください



SV_graphsリンクを
クリックします

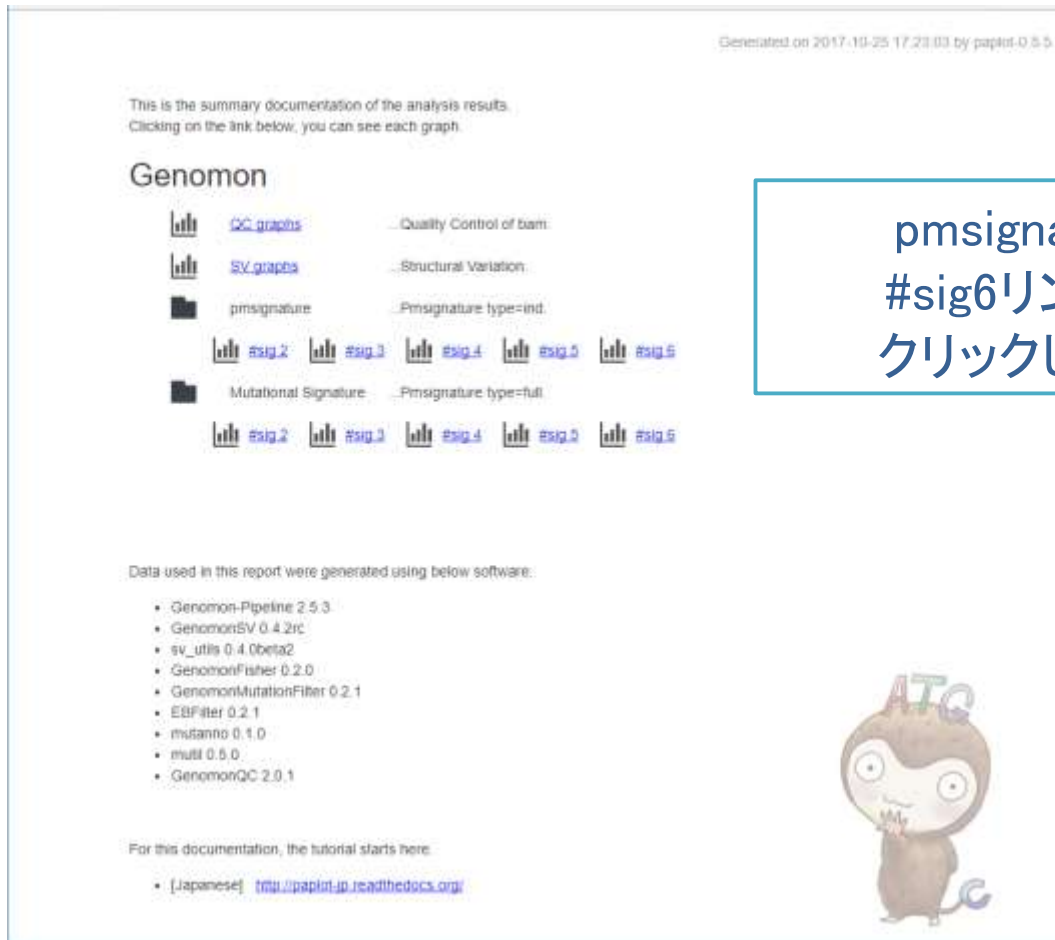
paplotでSVを確認する

Sample 5929の結果が表示されます。



paplotでSignature結果を確認する

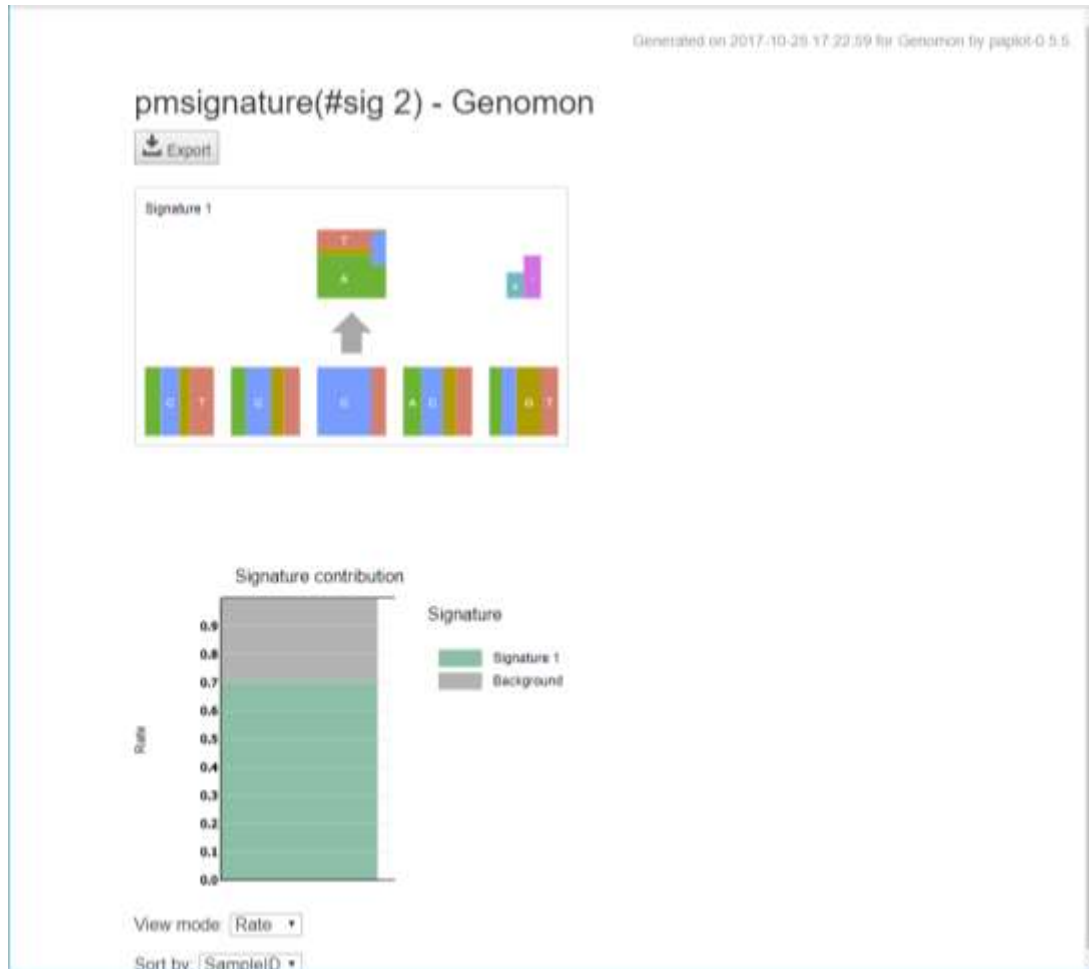
ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリックしてください



pmsignature
#sig6リンクを
クリックします

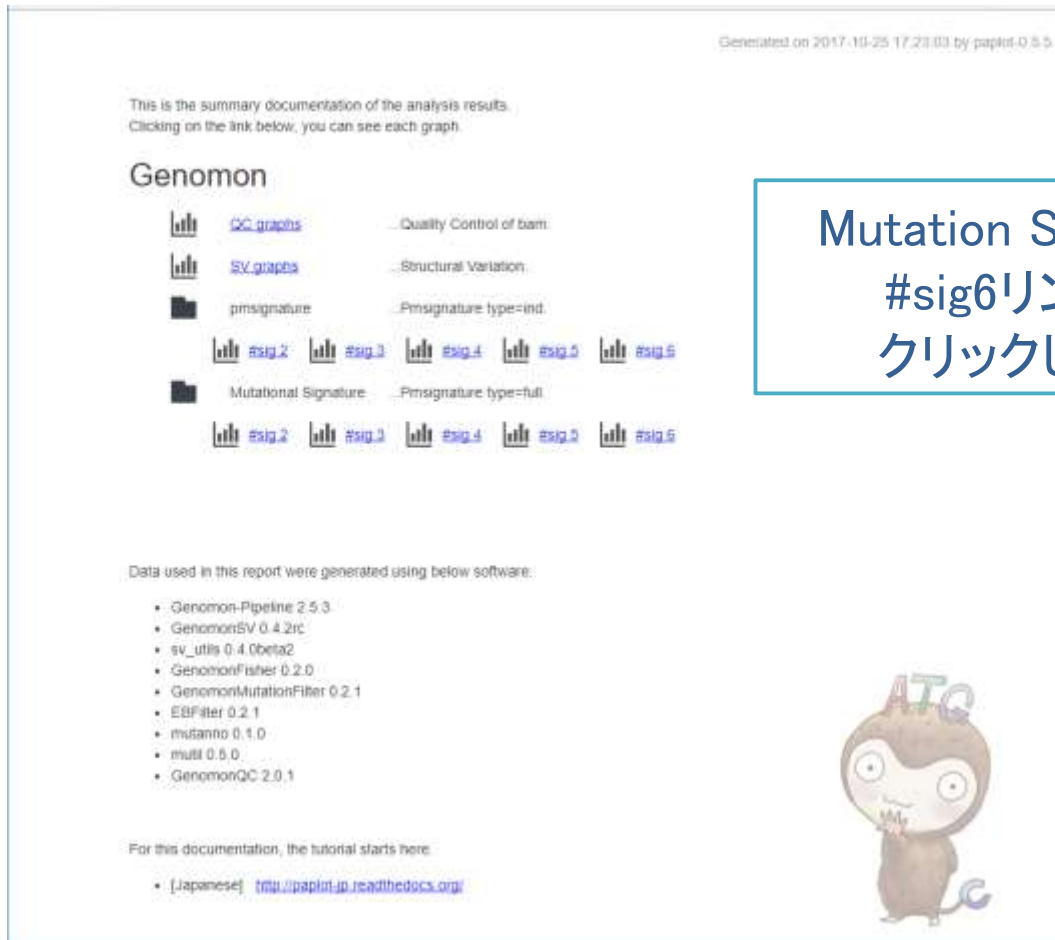
paplotでSignatureを確認する

test5929_exome(サンプルシート単位)の結果が表示されます。



paplotでSignature結果を確認する

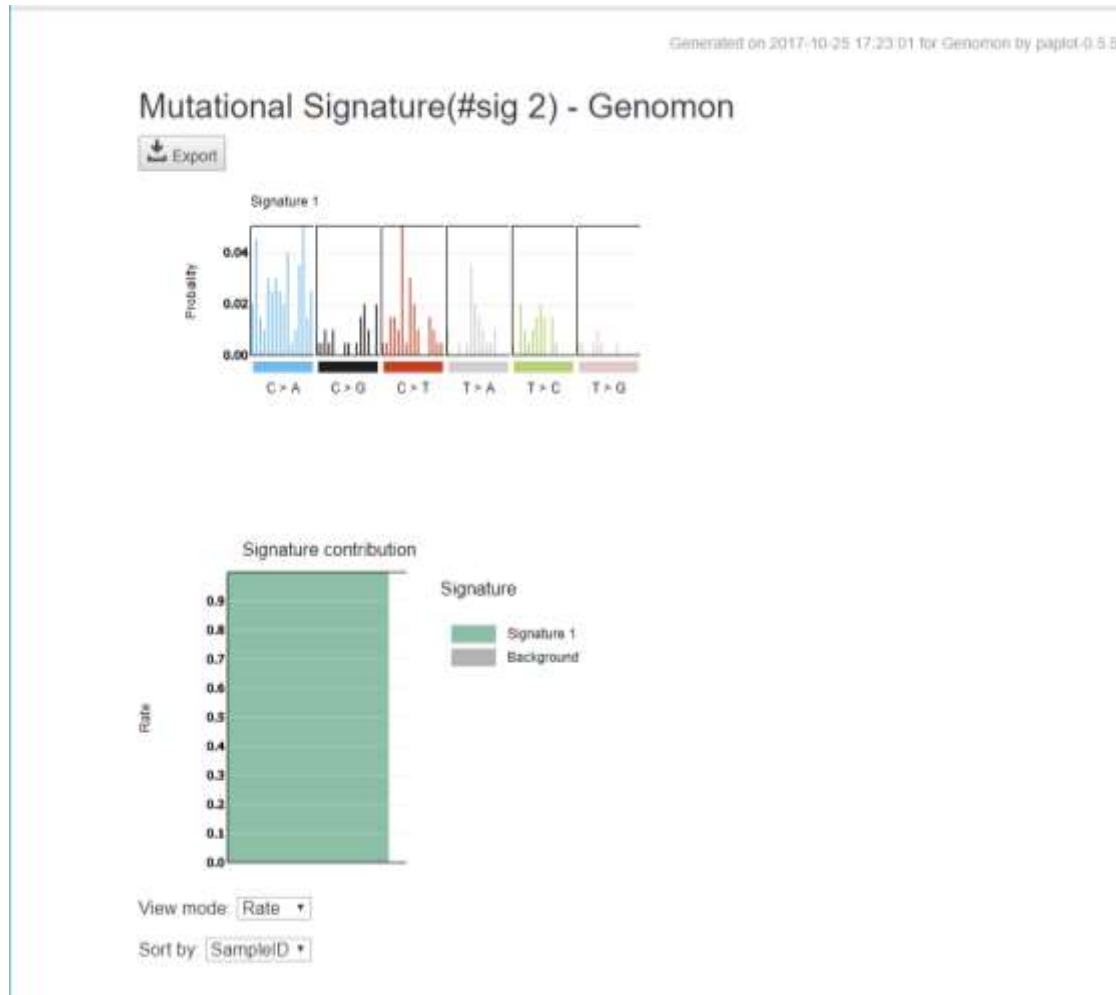
ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリックしてください



Mutation Signature
#sig6リンクを
クリックします

paplotでSignatureを確認する

test5929_exome(サンプルシート単位)の結果が表示されます。



本発表の流れ

- GenomonDNA解析の実行方法について説明
- DNA解析を実行してみましよう！
- DNA解析結果を確認しましょう。
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
 - **Genomon実行コマンド**について説明
 - サンプル設定ファイルの解説
 - パイプライン設定ファイルの解説
- RNA解析を実行してみましよう！（実習）
- RNA解析結果を確認しましょう。
- paplotの説明（RNA）

Genomonの実行 HGCスパコン用

```
$ bash /share/pub/genomon/genomon_pipeline-  
2.6.3/genomon_script/genomon_pipeline_HGC.sh (解析タイプ) (サンプル設定ファ  
イル) (出力ルートディレクトリ) (パイプライン設定ファイル)
```

引数1：解析タイプ

DNA解析をする場合は, `dna` を, RNA解析をする場合は, `rna` を指定します。

引数2：サンプル設定ファイル

解析対象のファイル (FASTQやBAM) のパスなどを記載したファイルを作成し, そのファイル名を指定します。

今日の実習ではこちらで用意したサンプル
MCF-7の解析をしていただきます。

引数3：出力ルートディレクトリ

出力ルートディレクトリを指定します。結果がこのディレクトリ以下に出力されます。

引数4：パイプライン設定ファイル

RNA解析用のパイプライン設定ファイルを指定します。

HGCスパコンに既に用意されているパイプ
ライン設定ファイルをコピーして使用します。

MCF-7サンプル設定ファイルの解説

[fastq]

MCF-7_test,/path/to/tumor/sequence1.fastq,/path/to/tumor/sequence2.fastq

[fusion]

MCF-7_test,None

[expression]

MCF-7_test

[intron_retention]

MCF-7_test

[qc]

MCF-7_test

[fastq]

(サンプル名), (FASTQ1ファイルパス), (FASTQ2ファイルパス)

サンプル名ですが、アルファベット、数値、ハイフン、アンダーバーは使用できます。あまり特殊な文字を使用しないでください。

入力情報タグで指定したサンプル名を使用して、解析情報タグを作成します。

[fusion]

(Tumorサンプル名), (コントロールパネル名)

[expression]

(Tumorサンプル名)

[intron_retention]

(Tumorサンプル名)

[qc]

(サンプル名)

[controlpanel]

(コントロールパネル名), (Normalサンプル名1), (Normalサンプル名2) ... (Normalサンプル名N)

*controlpanelがない場合は、Noneを指定します。

サンプル設定ファイルの項目の説明

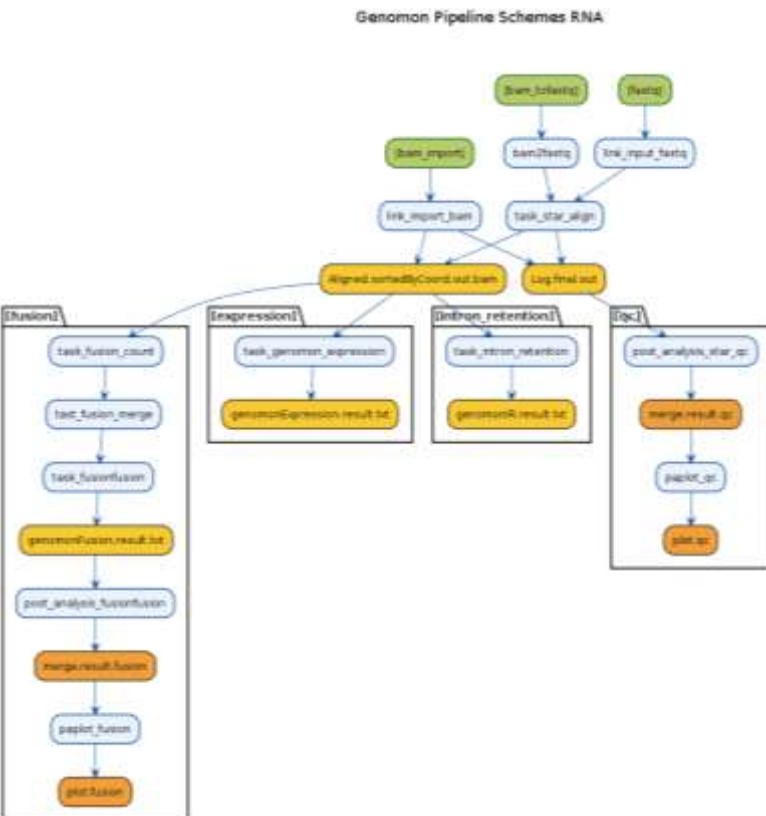
• 入力情報タグ

入力ファイルの情報を記載します

[fastq] FASTQファイルを入力として解析します

[bam_tofastq] BAMを入力してFASTQファイルに戻してから解析します

[bam_import] 入力したBAMを再マッピングせずに解析を実行します (Genomon2でアライメントしたBAMファイルが対象です)



• 解析情報タグ

解析するための情報を記載します

[fusion] 融合遺伝子を検出します

[expression] 発現量を計算します

[intron_retention] Intron_retentionを検出します

[qc] BAMのQuality Controlを出力します

[controlpanel] コントロールパネルを指定します

パイプライン設定ファイル RNAの解説 (advanced)

[REFERENCE]

prepared reference fasta file

star_genome = /xxx/database/GRCh37.STAR-STAR_2.4.0k

...(省略)...

[REFERENCE]には

使用するリファレンスゲノムやデータベースが記載されています

[SOFTWARE]

prepared tools

samtools = /xxx/tools/samtools-1.2/samtools

STAR = /xxx/tools/STAR-STAR_2.4.0k/bin/Linux_x86_64/STAR

...(省略)...

[SOFTWARE]には

使用するソフトウェアのパスが記載されています

parameters for star alignment

[star_align]

qsub_option = -pe def_slot 6 -l s_vmem=5.3G,mem_req=5.3G

star_params = --runThreadN 6 --outSAMstrandField intronMotif --outSAMunmapped Within --

alignMatesGapMax 500000 --alignIntronMax 500000 --outSJfilterOverhangMin 12 12 12 12 --

outSJfilterCountUniqueMin 1 1 1 1 --outSJfilterCountTotalMin 1 1 1 1 --chimSegmentMin 12 --

chimJunctionOverhangMin 12 --outSAMtype BAM Unsorted

samtools_sort_params = -@ 6 -m 3G

[star_align]などでは、

各ステージで使用するパラメータやメモリ量などを調整できます

HGCに用意されているパイプライン設定ファイルは、基本的なRNA解析に適したパラメータの値となっておりますので、まずは、値を変更しないで動かして、出力結果をみて調整しましょう。

本発表の流れ

- GenomonDNA解析の実行方法について説明
- DNA解析を実行してみましよう！
- DNA解析結果を確認しましょう。
- paplotの説明
- GenomonRNA解析の実行方法について説明
- **RNA解析を実行してみましよう！**
 - **GenomonでRNA解析（実習）**
- RNA解析結果を確認しましょう。

実習！

#1 Tera TermでHGCログインノードにログインする (windowsのデスクトップにTeraTermがあります)

#2 HGCスパコンにログインする
\$ **qlogin**

#3 こちらで用意したサンプル設定ファイルを、作成したディレクトリにコピーする
\$ **cp /share/pub/genomon/genomon_pipeline-2.6.3/sample_sheet/test_rna/MCF-7_sample.csv Genomon2_6_3/config**

#4 こちらで用意したパイプライン設定ファイル(RNA用)を、作成したディレクトリにコピーする
\$ **cp /share/pub/genomon/genomon_pipeline-2.6.3/genomon_conf/rna_genomon.cfg Genomon2_6_3/config**

#5 Genomon2の実行 (以下、2行に分かれていますが、一行のコマンドです。一度に実行してください)
\$ **bash /share/pub/genomon/genomon_pipeline-2.6.3/genomon_script/genomon_pipeline_HGC.sh rna**
Genomon2_6_3/config/MCF-7_sample.csv Genomon2_6_3/testMCF-7_rna Genomon2_6_3/config/rna_genomon.cfg

#6 Jobが実際に動いているか確認する
\$ **qstat** (watch qstat コマンドでモニタリングできます。完了したらCtrl+cでキャンセルしましょう)

Jobが完了したら次のページへ

#8 WinSCPを使用して以下のファイルを任意のディレクトリにダウンロードしましょう

QC結果

[Genomon2_6_3/testMCF-7_rna/post_analysis/MCF-7_sample/merge_starqc.txt](#)

fusion検出結果

[Genomon2_6_3/testMCF-7_rna/post_analysis/MCF-7_sample/merge_fusionfusion.txt](#)

[Genomon2_6_3/testMCF-7_rna/post_analysis/MCF-7_sample/merge_fusionfusion_filt.txt](#)

発現量計算結果 (post_analysisには結果が出力されません)

[Genomon2_6_3/testMCF-7_rna/expression/MCF-7_test/MCF-7_test.genomonExpression.result.txt](#)

intron_retention検出結果 (post_analysisには結果が出力されません)

[Genomon2_6_3/testMCF-7_rna/intron_retention/MCF-7_test/MCF-7_test.genomonIR.result.txt](#)

paplot結果 (ディレクトリごとダウンロード)

[Genomon2_6_3/testMCF-7_rna/paplot](#)

#9 QC結果、fusion検出結果、発現量計算結果、intron_retention検出結果のテキストファイル(タブ区切りになってます)を、OpenOfficeやExcelなどのソフトで表示する

#10 [paplot/sample_config_RNA/index.html](#)をダブルクリックして表示する

本発表の流れ

- GenomonDNA解析の実行方法について説明
- DNA解析を実行してみましよう！
- DNA解析結果を確認しましょう。
- paplotの説明（DNA）
- GenomonRNA解析の実行方法について説明
- RNA解析を実行してみましよう！
- **RNA解析結果を確認しましょう。**
- paplotの説明（RNA）

Fusion検出の結果を確認する

- Fusion検出

merge_fusionfusion.filt.txt

– SV検出結果. 適切なフィルタ済み

merge_fusionfusion.txt

–SV検出結果.

(こちらはadvanced user向け)

フィルタ済みのファイルもありますが、フィルタが解析サンプルに対して 適切かある程度評価する必要があります。

result.txt から result.filt.txt は どのようなフィルタを行っているか

`${sample}. Chimeric.out.sam`



①fusionの候補を検出する

`${sample}. genomonFusion.result.txt`



②おすすめフィルタでfusionの候補を絞る

`${sample}. genomonFusion.result.filt.txt`

- ・あるはずの変異がresult.filt.txtにない！

→まずは、result.txtに変異候補がないか確認しましょう。

- ・result.txtにあった

→result.filt.txtに出力されない原因を確認しましょう（Normalに変異がはいっているため検出されないとか）

パイプライン設定ファイルを変更して ②のおすすめフィルタを変更する

```
# おすすめフィルタのコマンド  
/path/to/fusion_utils filt ¥  
$input.txt $output.txt ¥  
{filt_params}
```

[fusionfusion]

filt_params = --filter_same_gene -grc

filter_same_gene

1st breakpointと 2nd breakpointのGene名が同じ場合

grc

chromosome nameにchrがついていない場合は指定する

fusion-fusionの結果をフィルタする

- ChromosomeがGL00---(Scaffold)、hs37d5(デコイ)の場合はほとんど偽陽性なのでフィルタする.

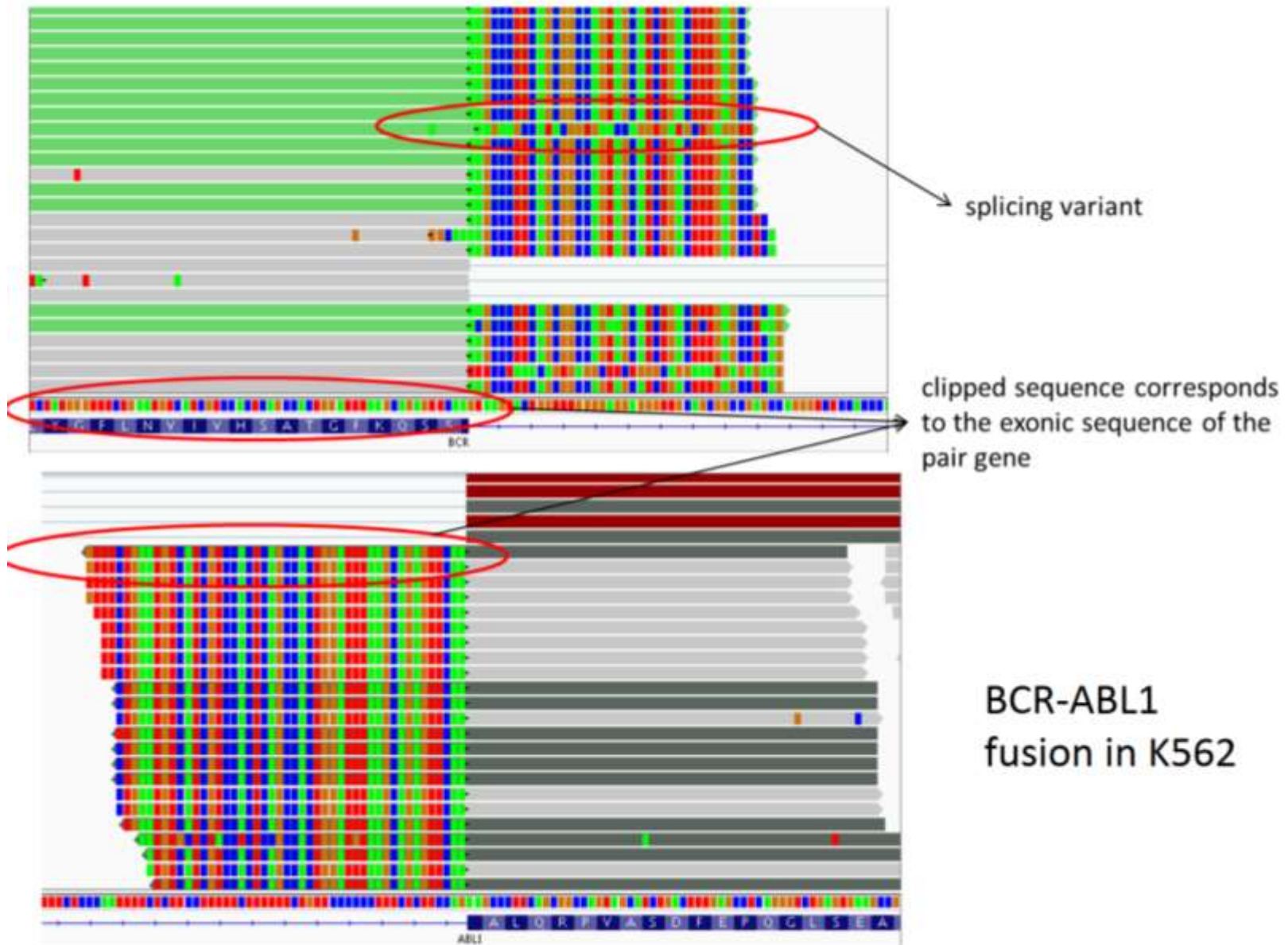
(fusion_utils filtでフィルタされます。)

- exon-intron junctionにstart, endがついていたら、確からしさが大幅にアップする。ついていなくても本物であることもある。

- supportingリードは3～5本くらいが閾値となる。

(fusion_utils filtでは3本でフィルタしています。)

Fusionが本物かチェックする



発現量を確認する

- ・結果ファイル

`$sample.genomonExpression.result.txt`

Gene名と発現量が記載されたファイルが出力されます.

Intron Retentionの候補を確認する

- ・結果ファイル

`$sample.genomonIR.result.txt`

Intron Retentionの候補を確認する.

BAM QCの結果を確認する

DNAのQCの計算方法が異なります。

RNAのQCはSTARが出力した '\$sample.Log.final.out' を使用します。

結果ファイル

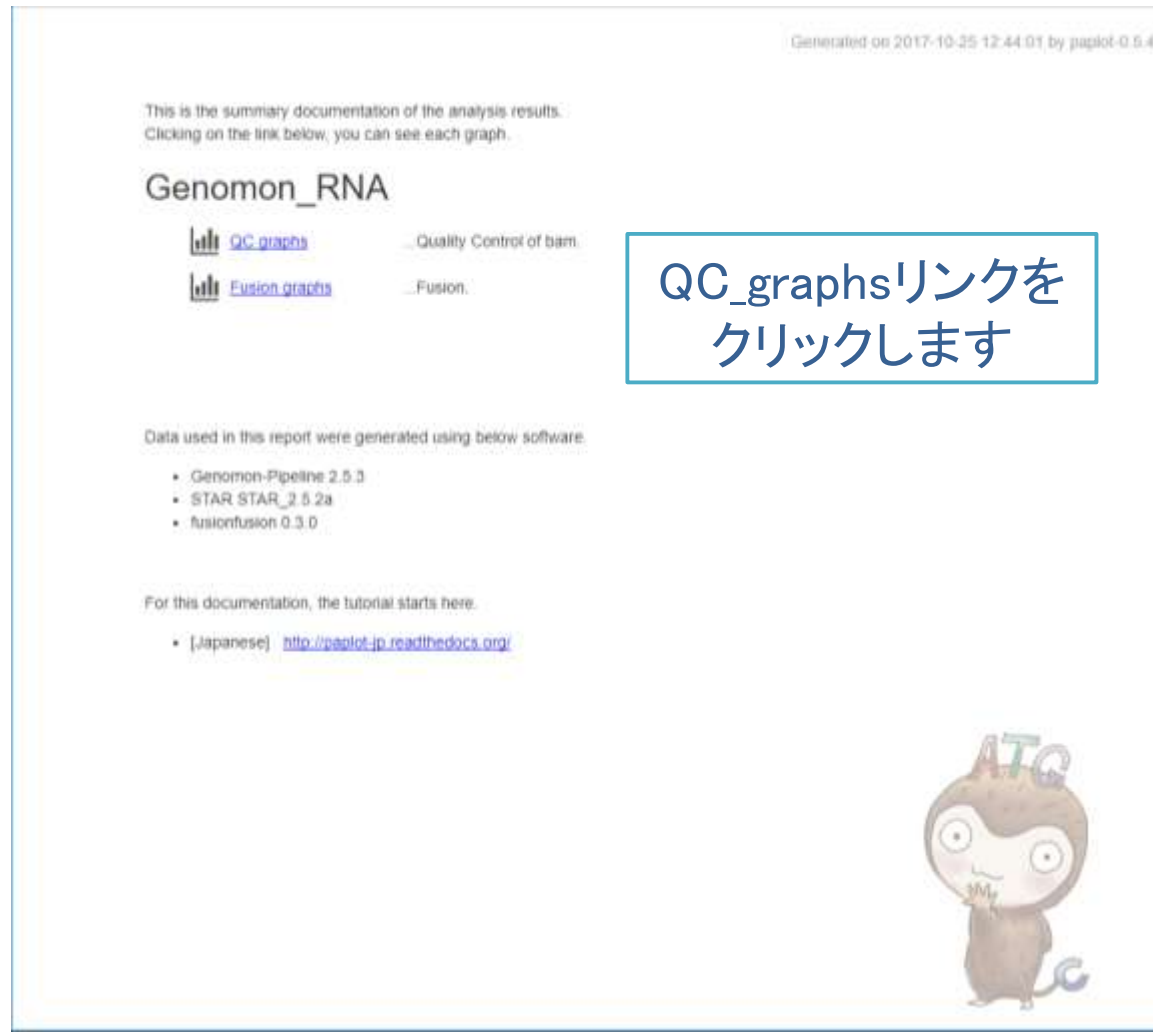
post_analysis/\$サンプル設定ファイル名/merge_starqc.txt

確認する項目：

Uniquely mapped reads %,

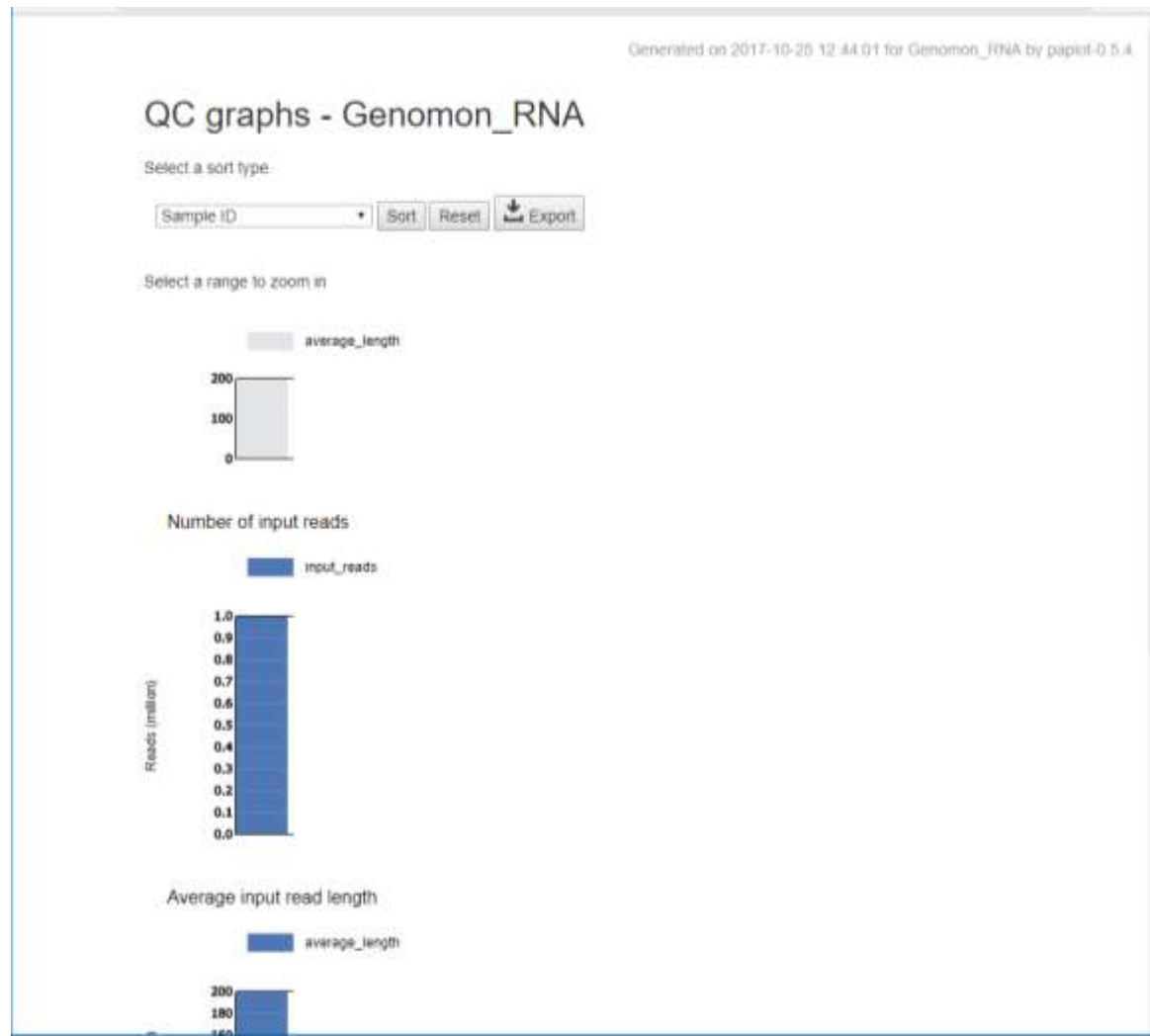
paplotでQCを確認する

ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリックしてください



paplotでQCを確認する

Sample MCF-7の結果が表示されます.





paplotでfusion結果を確認する

ダウンロードしていただいたGenomon実行結果の
paplot/5929_sample/index.htmlをダブルクリップしてください

Generated on 2017-10-25 12:44:01 by paplot-0.5.4

This is the summary documentation of the analysis results.
Clicking on the link below, you can see each graph.

Genomon_RNA

| | |
|---|------------------------|
|  QC graphs | Quality Control of bam |
|  Fusion graphs | Fusion |


Data used in this report were generated using below software:

- Genomon-Pipeline 2.5.3
- STAR STAR_2.5.2a
- fusionfusion 0.3.0

For this documentation, the tutorial starts here.

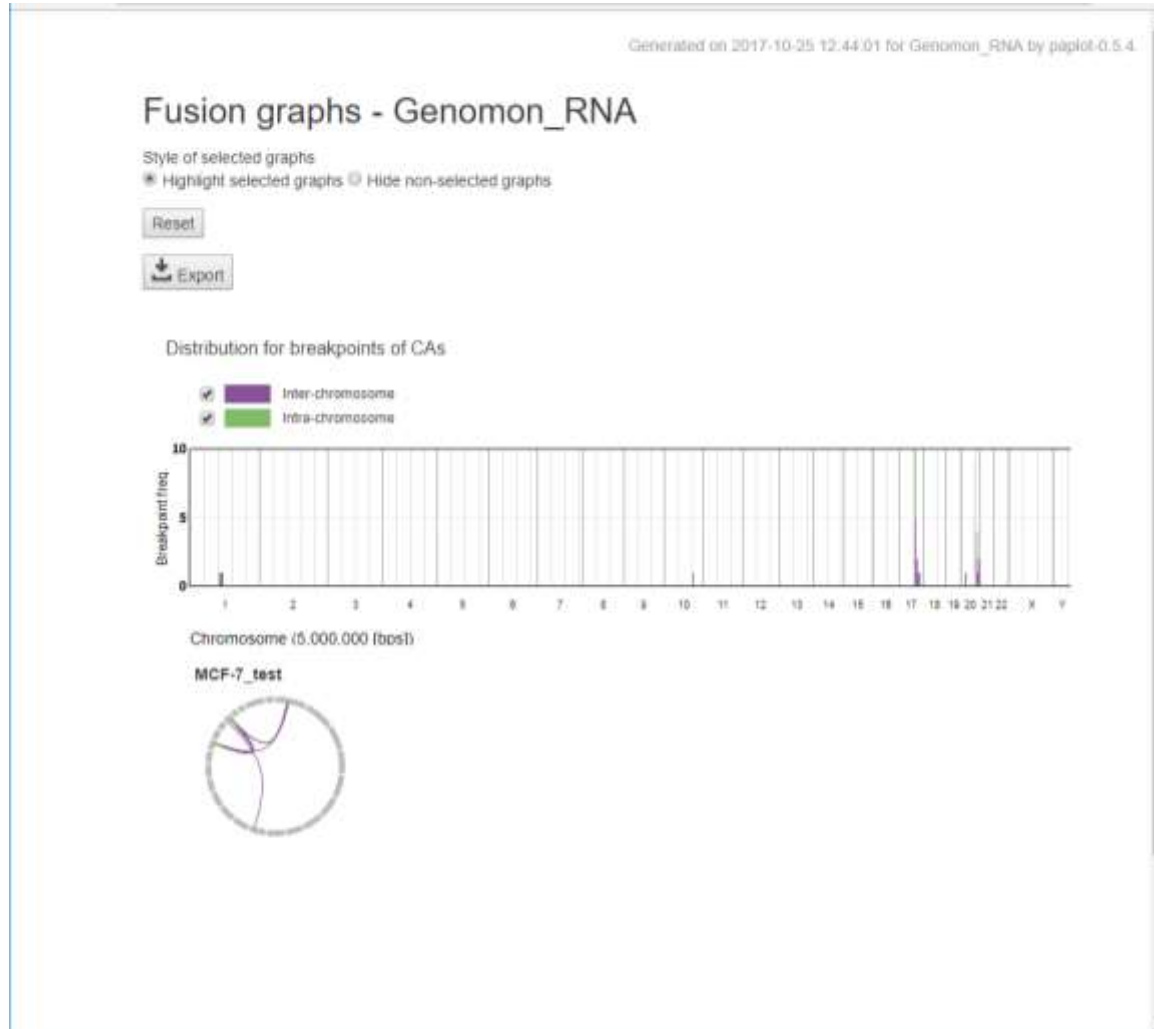
- [Japanese] <http://paplot.jp/readthedocs.org/>

Fusion_graphsリンクを
クリックします



paplotでfusionを確認する

Sample MCF-7の結果が表示されます。



paplotの説明

多検体で解析した結果を用意しました.

こちらのサイトを開いてください.

http://genomon-project.github.io/paplot/index_ja.html

引用

Genomon2 Tutorial (2016 年 5 月 24 日実施)した
「Genomon2 を利用したがんゲノム解析の実際」で吉田
先生が発表された資料の一部をお借りしました。

Shirokane3 Tutorial/Consultingのセミナーで使用了た
資料をお借りしました。