

統合データベース講習会: AJACS出島
2014年7月3日

次世代シーケンスデータの視覚化

九州大学生体防衛医学研究所
佐藤 哲也

<sato@bioreg.kyushu-u.ac.jp>

受講対象者

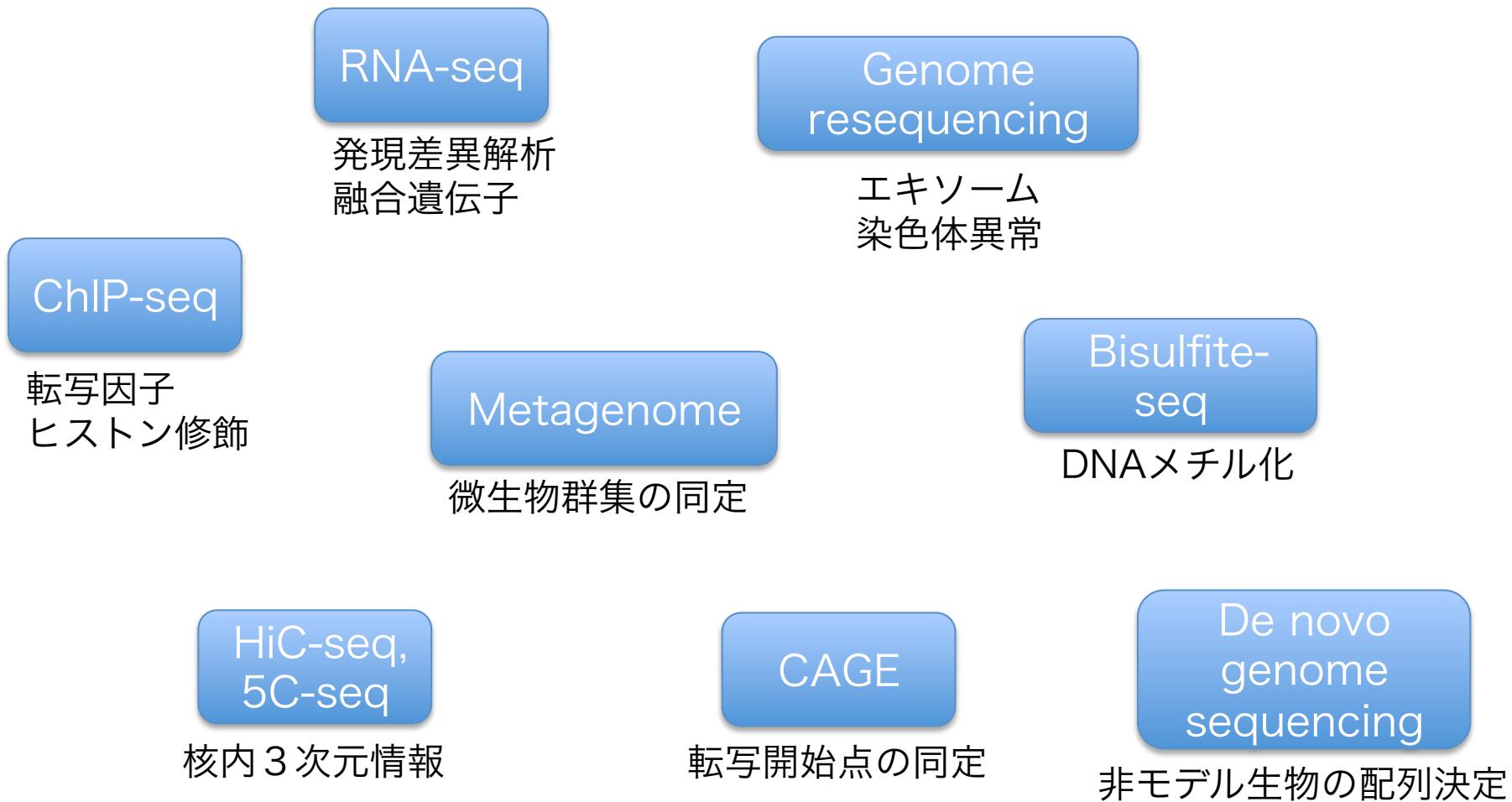
本日の講習会の内容は、次の方を対象としています。

- 次世代シーケンス解析に興味がある方、今後解析してみたい方
- 解析の経験はあるがデータ解析はバイオインフォマティシャンにおまかせしているのでデータの詳細についてはよく理解していないような方

本日の内容

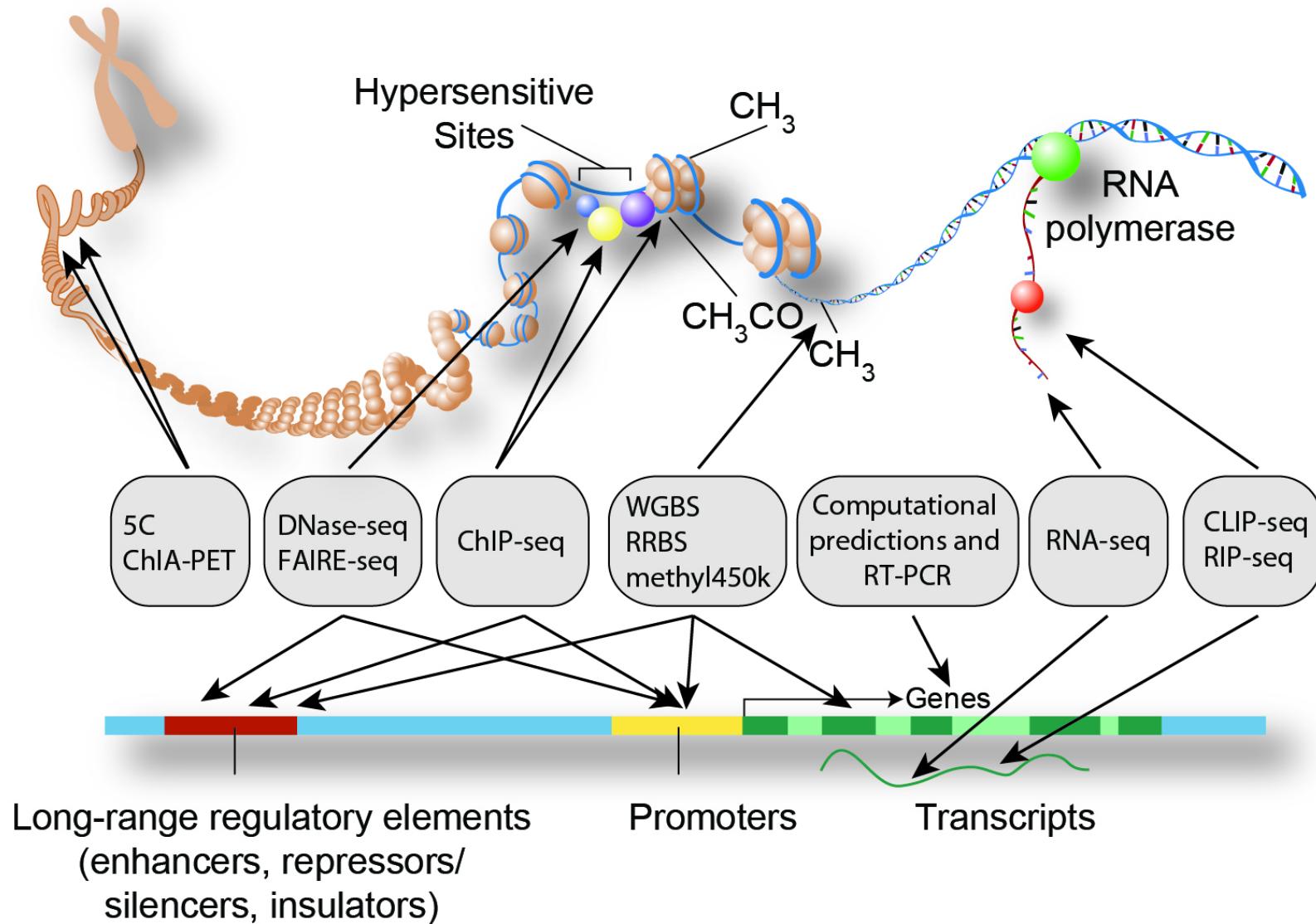
- 次世代シーケンス解析の概要
- 公共データの活用とデータの視覚化
- Integrative Genomics Viewer (IGV) の利用
- IGVを使った演習

次世代シーケンス (NGS) 解析

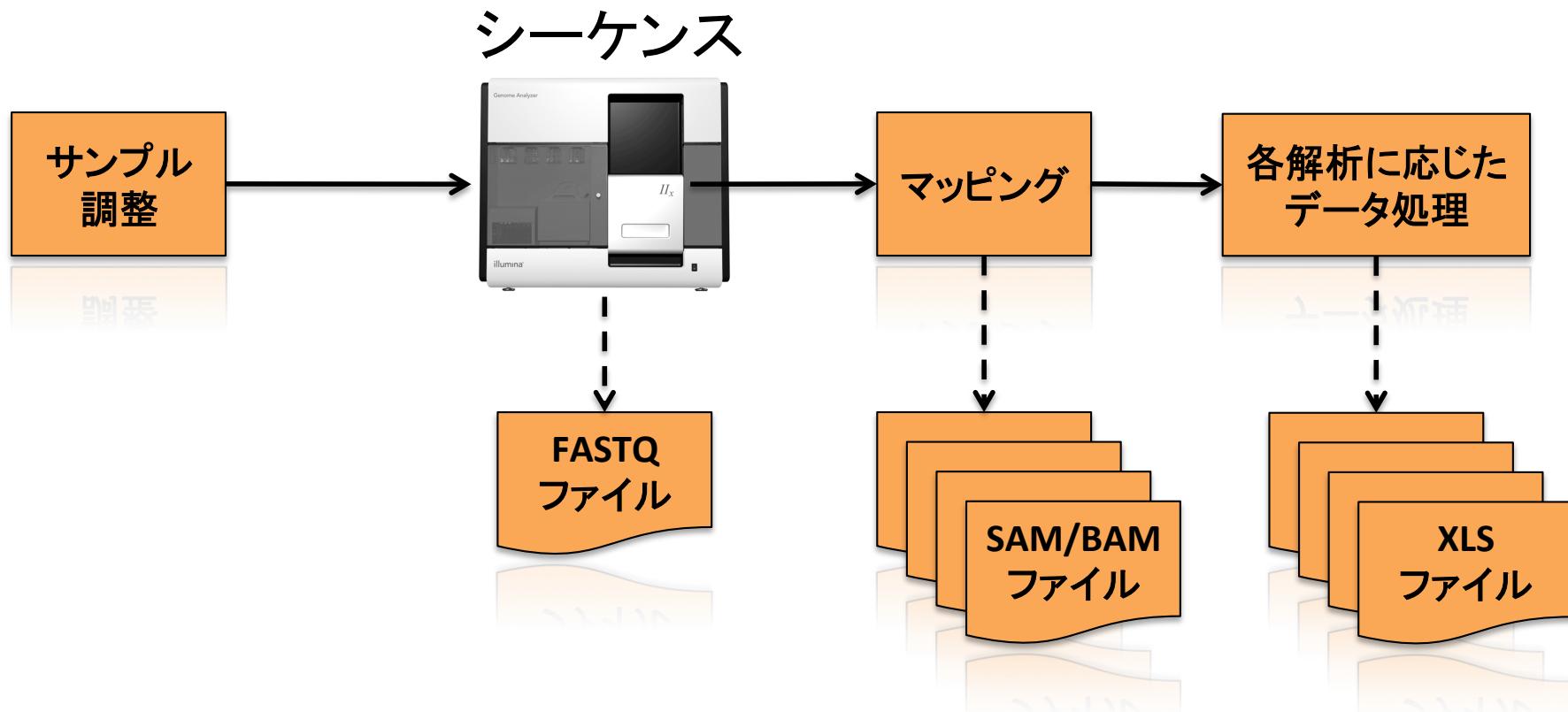


※本講習では、RNA-seq解析とChIP-seq解析のデータを対象とします。

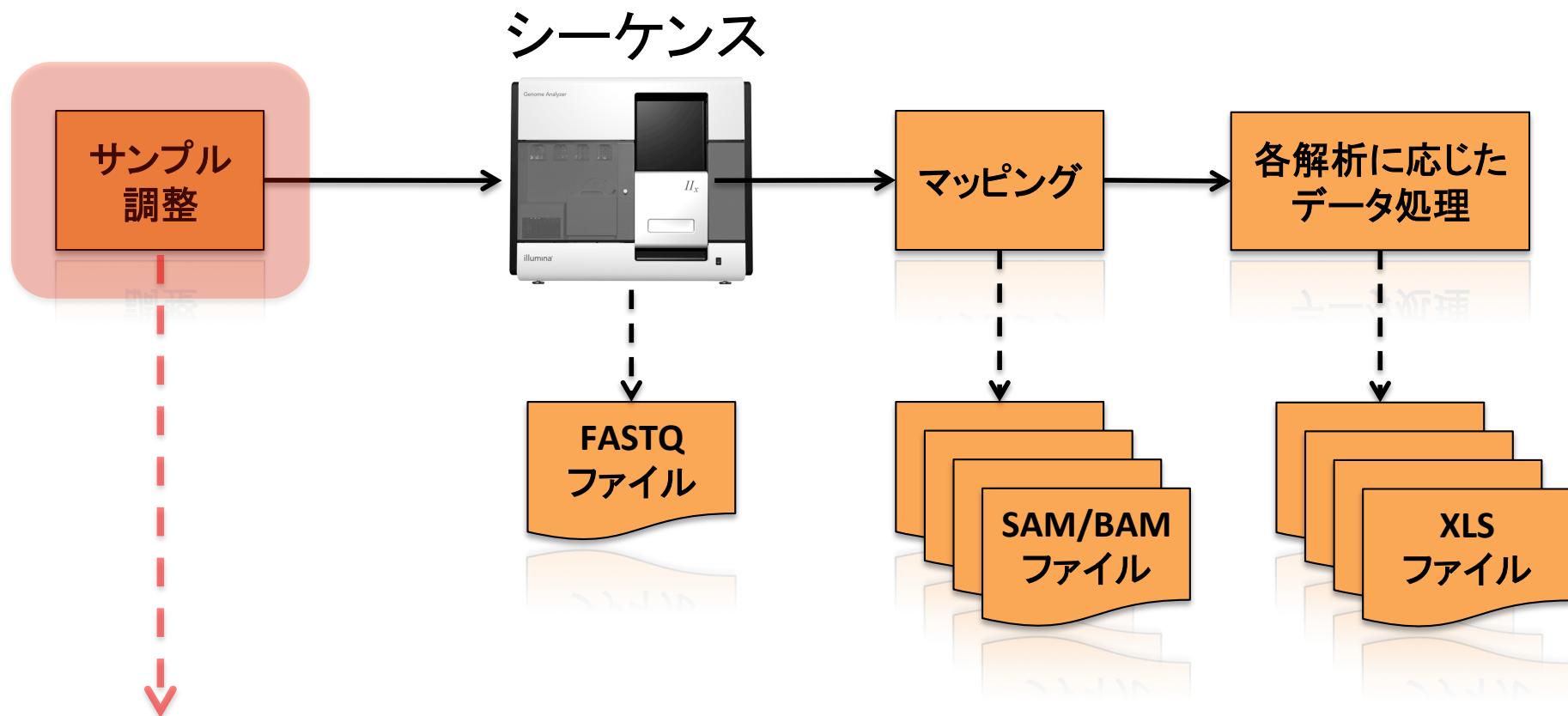
NGS解析の組み合わせ



次世代シーケンス解析概要



次世代シーケンス解析概要



解析の目的に従ってDNAを精製してサンプルを調整します。
(single-end read, paired-end read, strand-specific, ...)

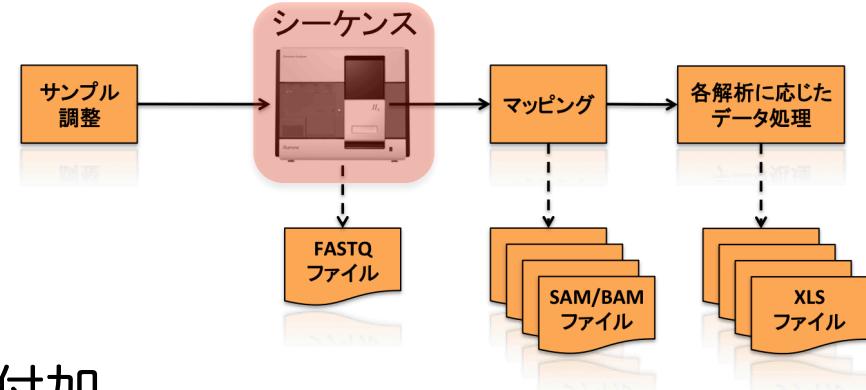
次世代シーケンサー技術

サンプルDNAの
断片化

アダプター配列の付加

アダプター配列を介して、
フローセル（基盤）にサンプルDNAを結合

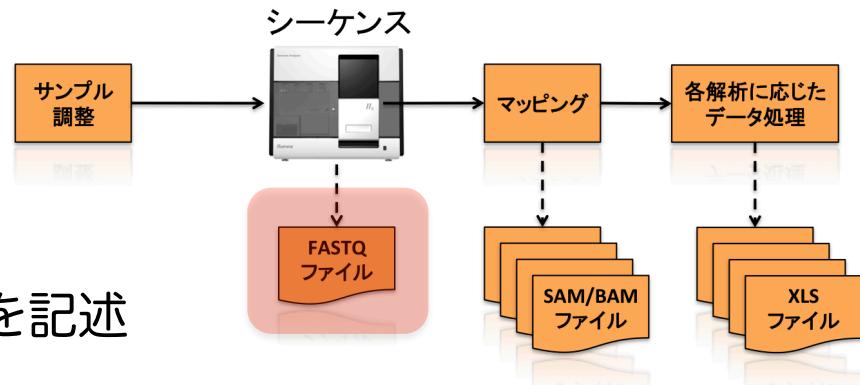
サンプルDNAの複製を一塩基づつ合成し、
蛍光強度の違いから塩基を同定
→FASTQファイル入手



FASTQ形式

塩基配列とクオリティスコアの両方の情報を1つのファイルに記述するファイル形式。

- 1行目 「@」で始まり、その後ろに配列IDを記述
 - 2行目 塩基配列を記載
 - 3行目 「+」を記載（加えて配列IDを記述することもある）
 - 4行目 2行目に記述した配列のクオリティスコアを記述

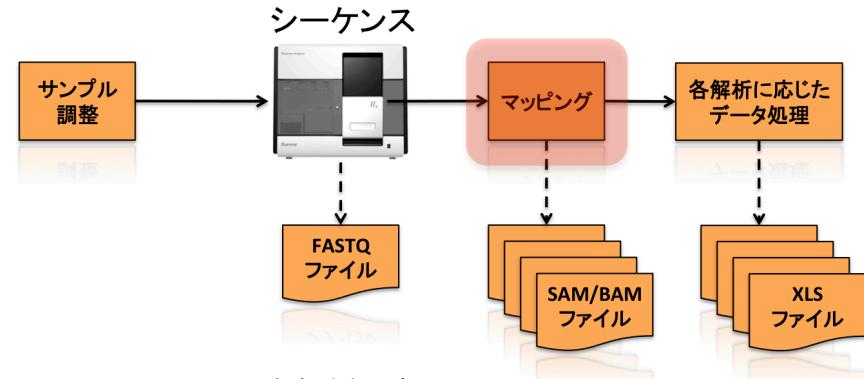


クオリティスコアは、ASCIIコードを用いて一文字で表現されます。
実際のスコアは次の通り。

33 33 34 37 37 37 37 39 39 39 39 39 38 41 41 41 40 41 ...

マッピング

マッピング用のプログラムは世界中の研究者によって開発されているが、その多くはコマンドラインで操作します。



MAC環境やLinux環境であれば、ローカルマシンで比較的容易にプログラムを実行できるのですが、Windows環境だと難しい。

そこで、どのようなマシン環境でも解析可能な方法として、GalaxyまたはDDBJの解析パイプラインのご利用をお勧めします。

Galaxy <https://usegalaxy.org/> or <http://galaxy.dbcls.jp/>

DDBJ <https://p.ddbj.nig.ac.jp/pipeline/Login.do>

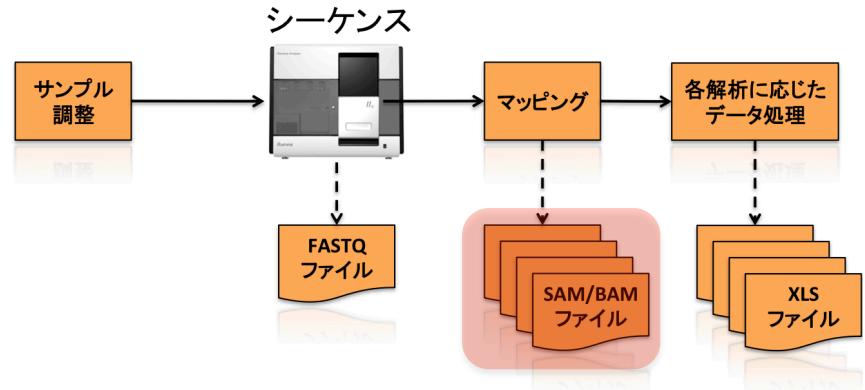
なお、本講習ではマッピングは行わず、マッピング済みデータを利用します。

本日の内容

- 次世代シーケンス解析の概要
- 公共データの活用とデータの視覚化
- Integrative Genomics Viewer (IGV) の利用
- IGVを使った演習

マッピングデータの活用

様々な国際プロジェクトでは、
マッピング後のデータを
公共データベースに登録しています。



- ENCODE (Encyclopedia of DNA Elements)
<http://genome.ucsc.edu/ENCODE/>
- NIH Roadmap Epigenomics Project
<http://www.roadmapepigenomics.org/>
- IHEC (International Human Epigenome Consortium)
<http://www.ihec-epigenomes.org/>

これらの登録データを活用することで、研究の可能性が広がることを期待します。

ENCODE The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

The NIH Roadmap Epigenomics Mapping Consortium

Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, Peggy J Farnham, Martin Hirst, Eric S Lander, Tarjei S Mikkelsen & James A Thomson

The NIH Roadmap Epigenomics Mapping Consortium aims to produce a public resource of epigenomic maps for stem cells and primary *ex vivo* tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease.

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 10 OCTOBER 2010

Welcome to IHEC

Annual Meeting

IHEC Datasets

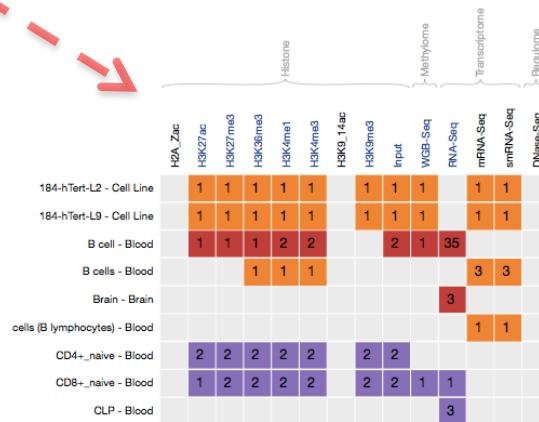
Lectures

Research

Why Epigenomics?

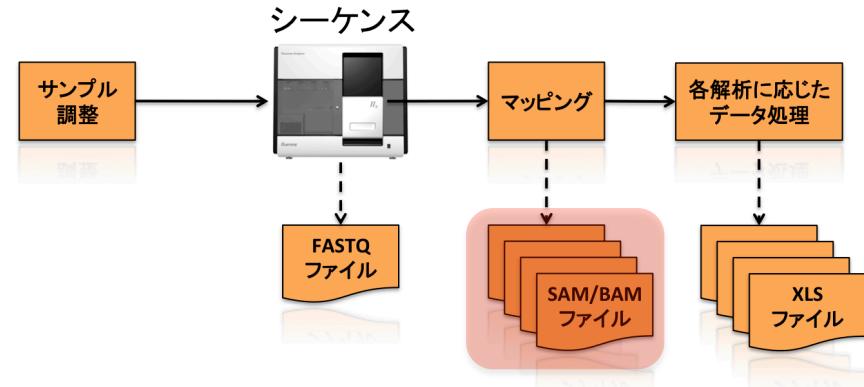
IHEC Data Portal

国際ヒトエピゲノムコンソーシアムThe International Human Epigenome Consortium (IHEC)は、様々な病気や生命現象に関わるヒトエピゲノムの情報を世界各国の研究者で協調・分担して解析し、いまだかつてない高精度のヒトエピゲノム地図をつくることを目的とした国際的な研究組織です。



マッピングデータの視覚化

マッピングデータを視覚化することで、データの解釈が容易になります。



これまで敷居が高かったデータベースを有効利用できるようになるかもしれません。

例えば、DRA (DDBJ) やSRA (NCBI) では、FASTQファイルの種類は検索できますが、どの遺伝子がどれくらい発現しているとか、ゲノム上のどの領域に転写因子が結合しているか等は検索できません。

本講習では、視覚化ツールの利用方法について説明したいと思います。

視覚化ツール

次世代シーケンス解析者用コミュニティーサイトのひとつにSEQanswers (<http://seqanswers.com/>) あります。このサイトには、次世代シーケンスデータ解析用ソフトウェアのリストがまとめられており、「Visualization」機能をもつものとして46個が登録されています（2014年6月28日現在）。

The screenshot shows a Wikipedia-style page titled "Visualization". The top navigation bar includes "Page", "Discussion", "Read", "View form", "View source", "View history", and a search bar. A sidebar on the left contains the SEQanswers logo, a "Forums" link, and sections for "wiki navigation" (Main page, Recent changes, Random page, Help), "Software", and "Toolbox". The main content area displays a table with 46 rows, each representing a software tool. The columns are: "Biological domain", "Bioinformatics method", "Input format", and "Output format". The table lists various tools like Anno-J, Avadis NGS, BamView, Biopieces, BLAST Ring Image Generator, Circos, CompreheNGSive, CummeRbund, DeepTools, and Galaxy, along with their specific features and supported formats.

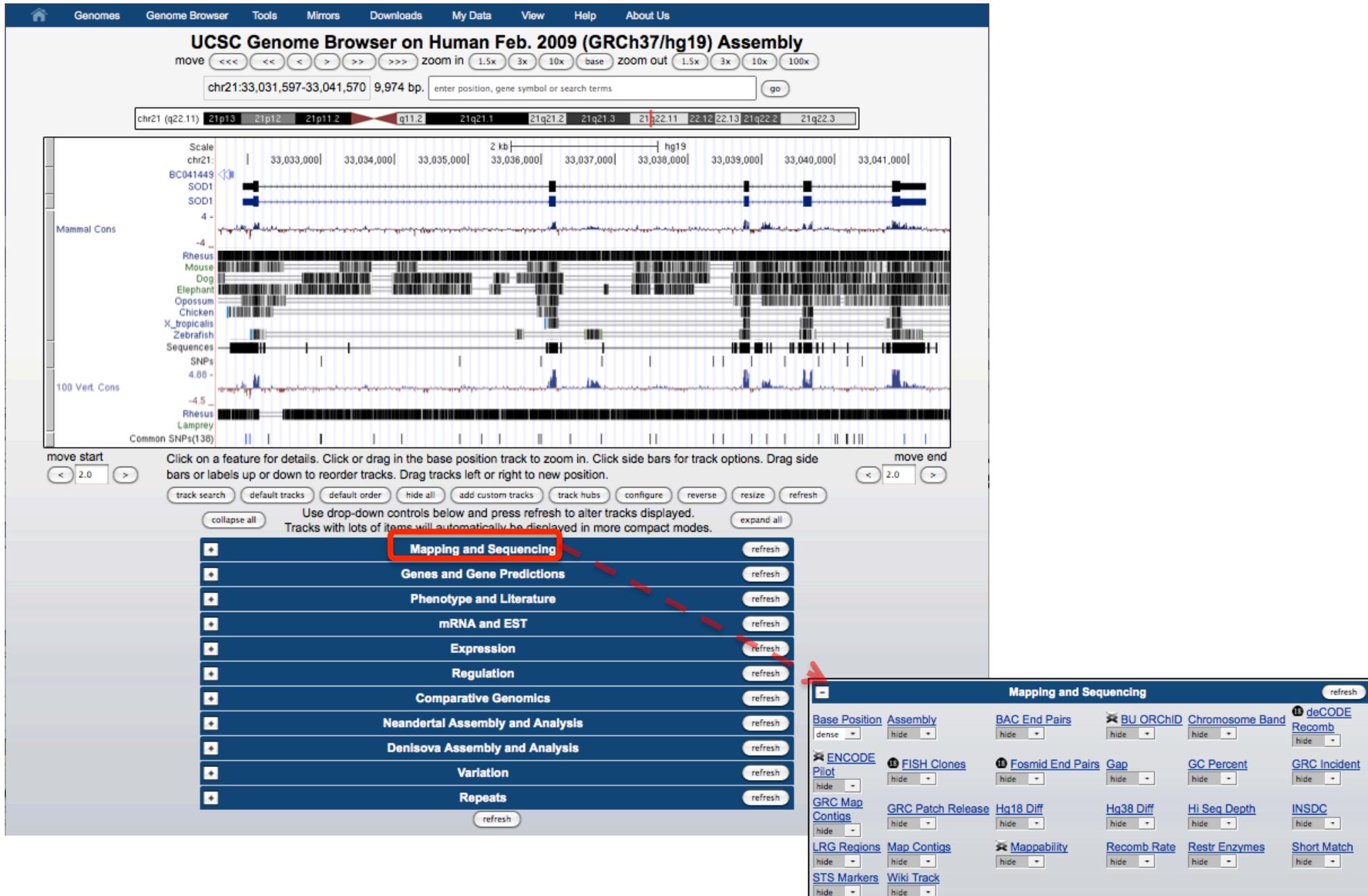
Biological domain	Bioinformatics method	Input format	Output format
Anno-J	Visualization		
Avadis NGS	ChIP-Seq DNA-Seq RNA-Seq Small RNA Pathway analysis	Alignment Quality Control Sequence analysis Visualization Biological Contextualization	SAM BAM BED ELAND FASTA FASTQ
BamView		Visualization	BAM
Biopieces	Genomics	Alignment Quality Control Sequence analysis Visualization	
BLAST Ring Image Generator	Comparative genomics	Visualization Assembly visualization	
Circos	Comparative genomics	Visualization	
CompreheNGSive	Next Generation Sequencing	Visualization	VCF SVG
CummeRbund	RNA-Seq Quantitation	Visualization	
DeepTools	Genomics ChIP-Seq	Normalization Visualization Conversion	Bed Bam Sam BEDGraph BedGraph BigWig Bed Bam Sam BEDGraph BedGraph BigWig
Galaxy	Comparative genomics Functional Genomics Whole Genome Resequencing Genomic Assembly Genomics	Alignment Assembly Quality Control Visualization	

<http://seqanswers.com/wiki/Visualization>

よく知られている視覚化ツール

- Integrative Genome Viewer (IGV)
- UCSC Genome Browser
- Ensembl Genome Browser
- Maqview
- SAMtools Text Alignment Viewer
- Tablet
- GenomeJack (三菱スペース・ソフトウェア株式会社)
- ZENBU (理研)
- ...

UCSC Genome Browser



本日の内容

- 次世代シーケンス解析の概要
- 公共データの活用とデータの視覚化
- Integrative Genomics Viewer (IGV) の利用
- IGVを使った演習

Integrative Genomics Viewer (IGV)とは？



James T. Robinson *et al.*, Integrative Genomic Viewer. *Nature Biotechnology* **29**, 24-26 (2011).

アレイデータや次世代シーケンスデータなどのゲノムデータを視覚化するためのツール。Broad Institute (米国) で開発。

【利点】 ゲノムワイドなデータや、様々な解析のデータが可視化できる

【欠点】 初期のデータ解析（マッピング）や、大量なデータの扱いには不向き

【データ解析】

- motif finder 塩基配列パターンが一致する領域を見つけることができる
- gene list view ある遺伝子セットのコーディング領域を一度にまとめて可視化

IGVの入手

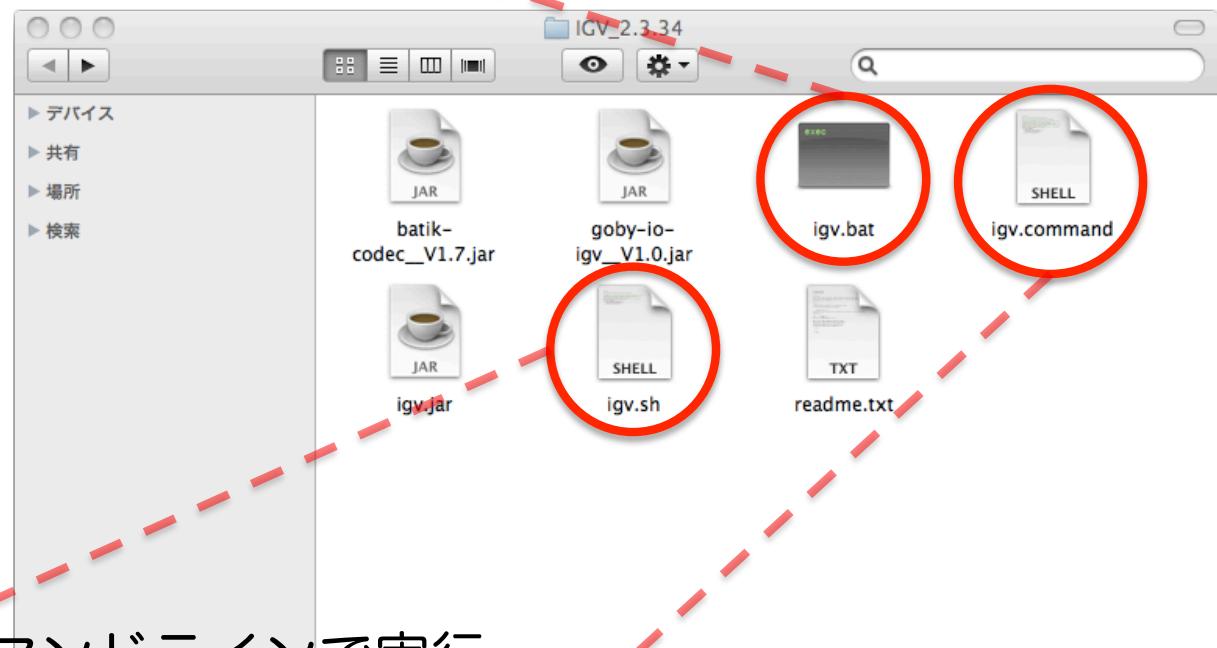
<http://www.broadinstitute.org/software/igv/home>

The screenshot shows the IGV website's home page. On the left, a sidebar contains links like 'Home', 'Downloads' (circled in red), 'Hosted Genomes', 'FAQ', 'User Guide', 'File Formats', 'Release Notes', 'Credits', and '@ Contact'. Below the sidebar is a search bar and the Broad Institute logo. The main content area features a large 'Integrative Genomics Viewer' title and a preview of the software's interface. It includes sections for 'What's New' (mentioning an ASHG annual meeting in October 2013 and the release of version 2.3 in April 2013), 'Citing IGV' (with a citation from Nature Biotechnology 2012), 'Overview' (describing IGV as a high-performance visualization tool), 'Funding' (listing funding from the National Cancer Institute, National Institute of General Medical Sciences, National Institutes of Health, and Starr Cancer Consortium), 'Downloads' (with a download link and a circled 'Log In' button), and logos for the National Cancer Institute, National Human Genome Research Institute, and GENOME SPACE.

The diagram illustrates the user flow for obtaining IGV. It starts with the 'IGV Registration' form, which requires Name, Email, and Organization, and includes Agree and Cancel buttons. This leads to the 'Log In' page, which prompts users to register if they haven't already. It has fields for 'email address' and a 'Login' button. Finally, it leads to the 'Downloads' section where users can choose between 'Download Mac App' (circled in red), 'Download Binary Distribution' (circled in red), or Java Webstart options for different memory requirements.

IGVの起動

Windows環境 「igv.bat」 ファイルをダブルクリック



MAC環境、Linux環境
「igv.sh」 ファイルをコマンドラインで実行

MAC環境 「igv.command」 ファイルをダブルクリック

IGV入力用ファイル形式

【次世代シーケンスデータ】

SAM, BAM, Bedgraph, bigBed, WIG, bigWig, GFF

BED, TDF, igv, narrowPeak, broadPeak, cufflinks出力ファイル

【IGVでgenomeファイル作成に必要なデータ】

FASTA, Cytoband, GTF

【ゲノムコピー数解析データ】

CBS, CN, MAF, SEG, SNP, VCF

【その他】

LOH, GCT, GISTIC, RES, Goby, PSL, genePred, MUT, GWAS, ...

SAM形式、BAM形式

【SAM形式】

Sequence Alignment/Mapの略。アライメントデータの様々な情報（マッピング座標、挿入、欠失、リード配列のクオリティーなど）を含むタブ切りテキスト形式のファイル。ファイルサイズは大きい。

HISEQ:189:D2ARKACXX:7:2215:3881:36075	99	chr1	10466	3	50M	=	10534	118	CCTCGCGGTACCCCTAGCCGCCGCTCGCCGGTCTGACCTGAGGAGA
HISEQ:189:D2ARKACXX:7:2215:3881:36075	147	chr1	10534	3	50M	=	10466	-118	AGTACCAACCAAATCTGTCAGAGGACAAACGCACCTCCGCCCTCGCGGTG
HISEQ:189:D2ARKACXX:8:1305:8270:59480	417	chr1	11128	1	50M	=	12134	1056	ACGGGTGAACATTCTGTAATCGAAAAGCAGGGATCGACGCCCTTGCTG
HISEQ:189:D2ARKACXX:8:1206:11763:46495	417	chr1	11134	1	50M	=	12056	972	GGAACTTCTACTAACCTGAAAAGCAGGGATCGACGCCCTTGCTGCGACG
HISEQ:189:D2ARKACXX:8:1216:4278:67073	419	chr1	11190	0	50M	=	11264	124	CTACAGGACCCGCTGCTCACGGTCTGTCAGGGCAGGGCCCCCTGCTGGC
HISEQ:189:D2ARKACXX:8:1216:4278:67073	339	chr1	11264	0	50M	=	11190	-124	TTGCTTAGAGTGTGGCCACCCGCCCTCTGCGCCGCCGGGACTCTGAGG
HISEQ:189:D2ARKACXX:8:1112:6609:50679	417	chr1	11297	0	50M	=	11432	185	CGCCGGGGCACTGCAGGCCCTCTGCTTACTGTATAGTGTGGCACGCC
HISEQ:189:D2ARKACXX:8:2207:2722:90338	161	chr1	11302	0	50M	=	11612	360	GGGCACTTGCAAGGCCCTCTGCTTACTGTATAGTGTGGCACGCCCTG
HISEQ:189:D2ARKACXX:8:1112:6609:50679	337	chr1	11432	0	50M	=	11297	-185	GCGGGGCCCTTGTCTAACAGTAGTGGCGGATTATAAGGAAACAACCG
HISEQ:189:D2ARKACXX:7:2210:17185:88513	417	chr1	11464	0	50M	=	11617	203	ATTATAGGGAAACACCGAGACATATGCTGTTGGCTCTAGTAGCTCT
HISEQ:189:D2ARKACXX:8:2204:10813:14534	419	chr1	11470	0	50M	=	11555	135	GGGAAACACCCGGACATATGCTGTTGGCTCTAGTAGACTCTAAATA
HISEQ:189:D2ARKACXX:8:2204:10813:14534	339	chr1	11555	0	50M	=	11470	-135	TTTAAATTGTTAACTGATTACCATCAGAAATTGACTGTTCTGTATCCCAC

CCCCFFAADDFFHDBGGIGIIIFPBHHGIJJ<FHIEBD9>CCEDDDAAA=@,59 AS:i:-6 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:26C23
DCC=B??DDEBCCCCEFFFDFGHHREC?EHEIHIHIIIGHDAFAFDDDDD@& AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:50 YT;
CCCCFDFHHHHHHHHJJ AS:i:-10 XN:i:0 XM:i:2 XO:i:0 XG:i:0 NM:i:2 MD:
CB#FFFFFHGHHHHHJJ AS:i:-10 XN:i:0 XM:i:2 XO:i:0 XG:i:0 NM:i:2 MD:
@@@DDDDDHHHHHGGIJJJI1G:CFHHDPEFD@BG?GHJIIIEHHFFB AS:i:-5 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:29G20
DCDCDCDDCDDBBBDAADB=DDIIIIIGHGH>G>HGDIIFGFHFDFDCC@ AS:i:-5 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:19G30
B@#FFFFFHGHGJIIJJ AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:50 YT;
CCCCFFFFHHHHHHJJ AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:50 YT;
IJIIJI1GCGJJJJIEJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ AS:i:-11 XN:i:0 XM:i:2 XO:i:0 XG:i:0 NM:i:2 MD:
CCCCFFFFFHHHHHJJ AS:i:-6 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:16G33
CCCCFFFFFHHHHHJJ AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:50 YT;
JIIJJ AS:i:-6 XN:i:0 XM:i:1 XO:i:0 XG:i:0 NM:i:1 MD:Z:10G39

(BAM形式)

SAMファイルのバイナリ（圧縮版）ファイル。IGVで利用する時は、indexファイルが必要。

BED形式、TDF形式

【BED形式】

ゲノム上の座標情報を表すタブ切りテキストファイル。1列目に染色体番号、2列目に開始座標、3列目に終了座標の記述が必要。4列目以降の使用は任意（4列目名称、5列目スコア、6列目ストランド…）。

chr12	63686	63761	42YU4AAXX_HWUSI-EAS627_1:2:13:1183:272/2	0	+
chr12	64305	64380	42YU4AAXX_HWUSI-EAS627_1:1:112:406:1542/1	0	+
chr12	72538	72613	42YU4AAXX_HWUSI-EAS627_1:2:64:1643:1418/1	0	+
chr12	72664	72739	42FD1AAXX_HWUSI-EAS627_1:3:117:1083:1246/1	0	+
chr12	72947	73022	42YU4AAXX_HWUSI-EAS627_1:2:87:656:1460/2	1	+

【TDF形式】

リードの集積情報を含むバイナリデータ。ファイルサイズは小さく扱いやすい。

WIG形式、bigWig形式

【WIG形式】

ゲノム上の各座標にマッピングされたリード数を記述している。1行目ではWIGファイル形式を指定し、2行目以降にリード数に関する値が記述される。また、この形式には、「variableStep」と「fixedStep」の2種類がある。

```
variableStep chrom=chr2  
300701 12.5  
300702 12.5  
300703 12.5  
300704 12.5  
300705 12.5
```

```
fixedStep chrom=chr3 start=400601 step=100  
11  
22  
33
```

【bigWig形式】

WIGファイルのバイナリ（圧縮版）ファイル。

ファイル形式とファイルサイズ

【実際に解析したRNA-seqデータの例】

50 bp×2 (paired-end read)で読まれたシーケンスデータ。トータルリード数は、64,491,380×2 リード。

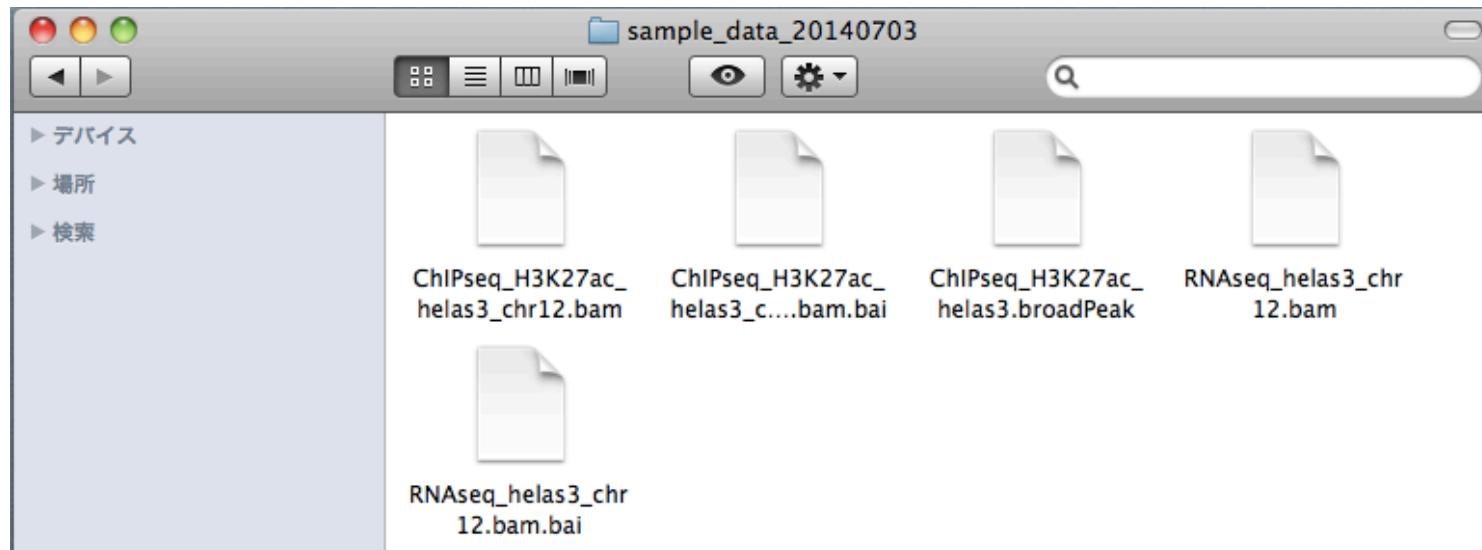
ファイル形式	ファイルサイズ
FASTQ	19,713MB
FASTQ (圧縮ファイル)	5,278MB
SAM	36,010MB
BAM	5,491MB
BED	9,536MB
WIG	452MB
bigWig	162MB
TDF	125MB

ファイルサイズは、TDF形式が最も小さい。

講習用データの説明

下記サイトから講習用データを入手してください。

http://bioinfo.sls.kyushu-u.ac.jp/sato/sample_data_20140703.tar.gz



ENCODEプロジェクトで解析されたデータ

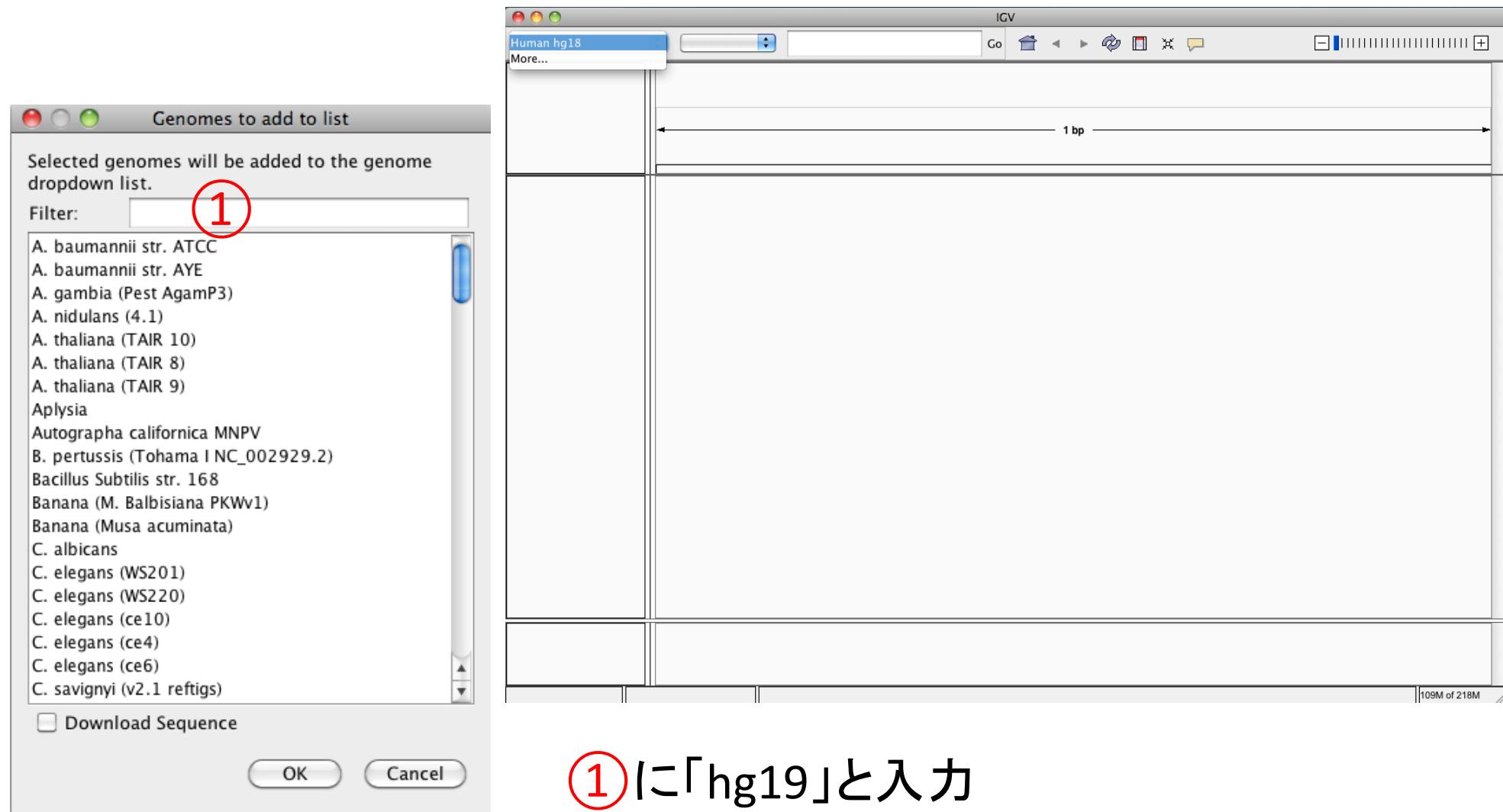
細胞の種類：HeLa S3

データの種類：RNA-seq、ChIP-seq

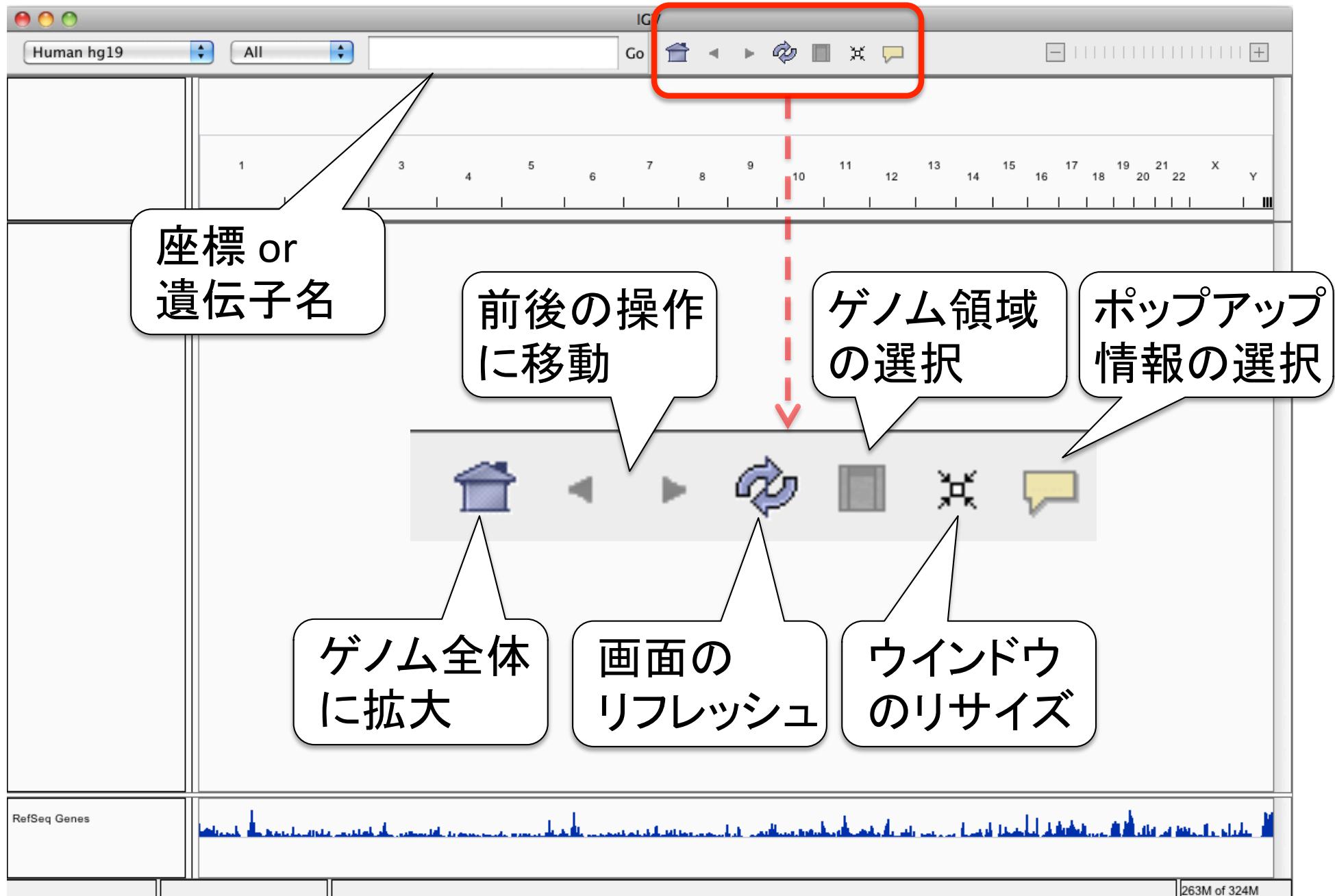
データサイズを小さくするために、12番染色体短腕のデータのみ

IGVの基本操作：ゲノムバージョン

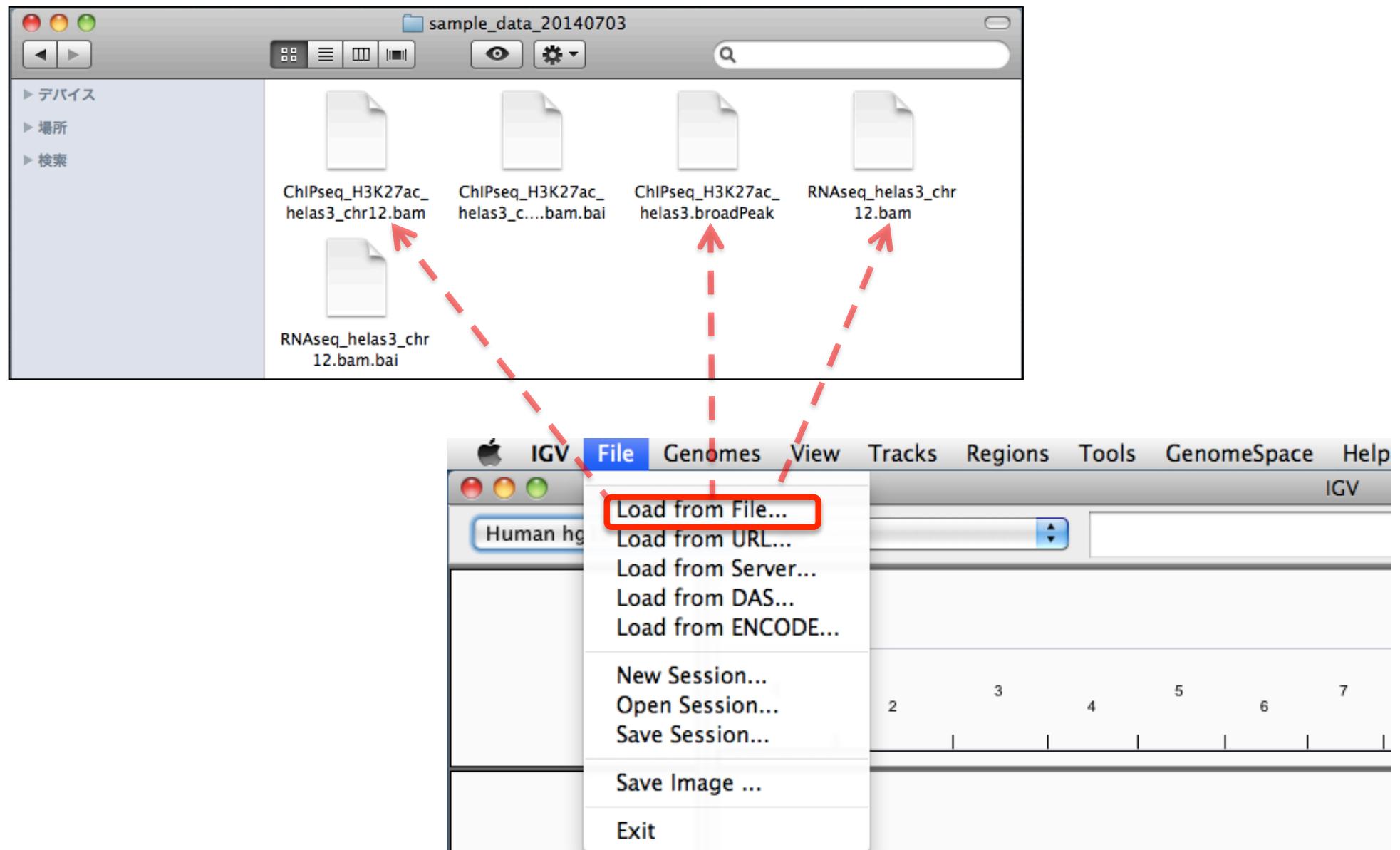
まず始めに、視覚化に用いるゲノムデータを選択します。



IGVの基本操作：アイコン



ファイルの読み込み

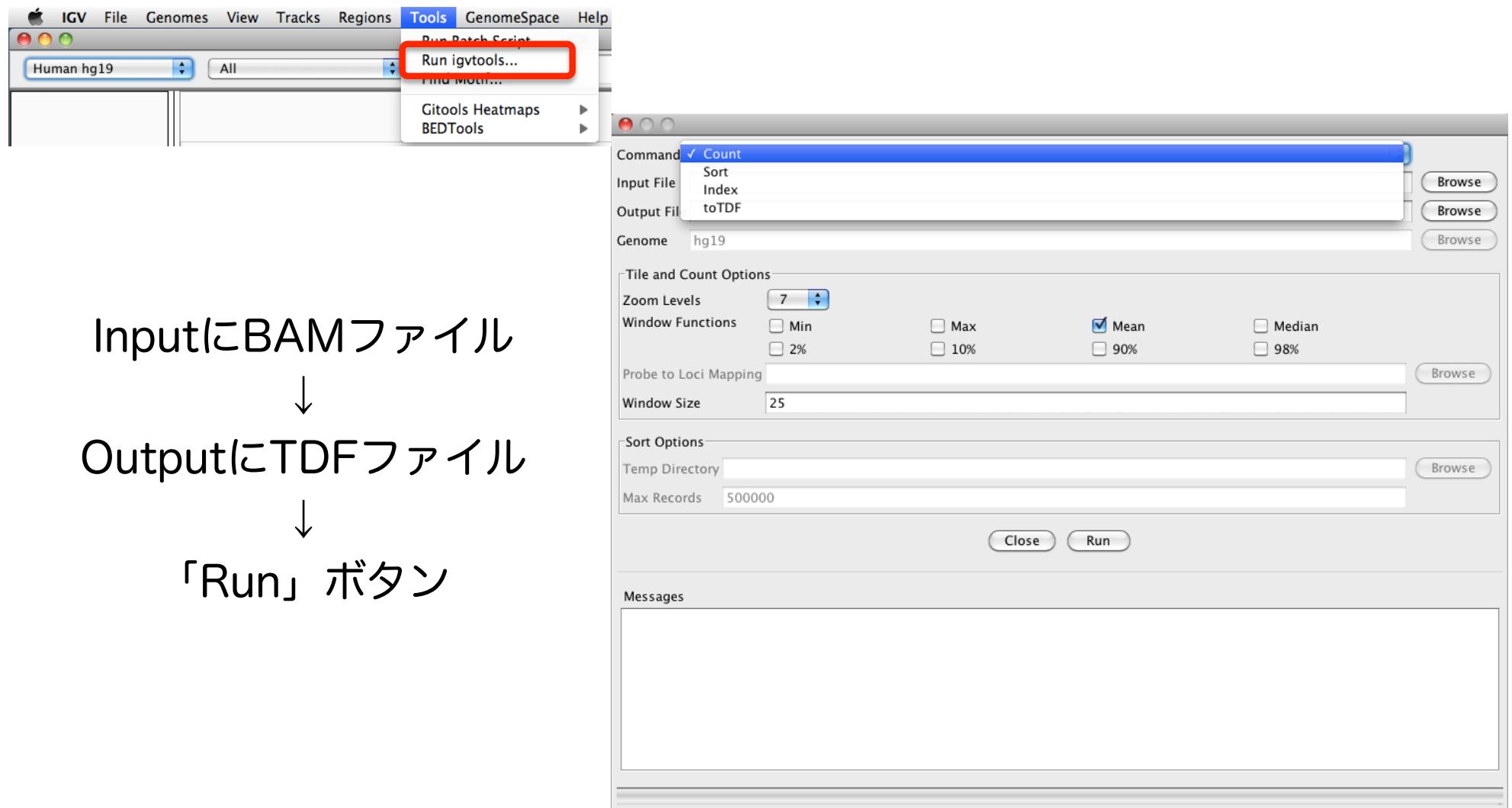


本日の内容

- 次世代シーケンス解析の概要
- 公共データの活用とデータの視覚化
- Integrative Genomics Viewer (IGV) の利用
- IGVを使った演習

演習1：TDFファイルを作成

IGVで読み込むことができるTDFファイルをIGVToolsで作成してください。



InputにBAMファイル



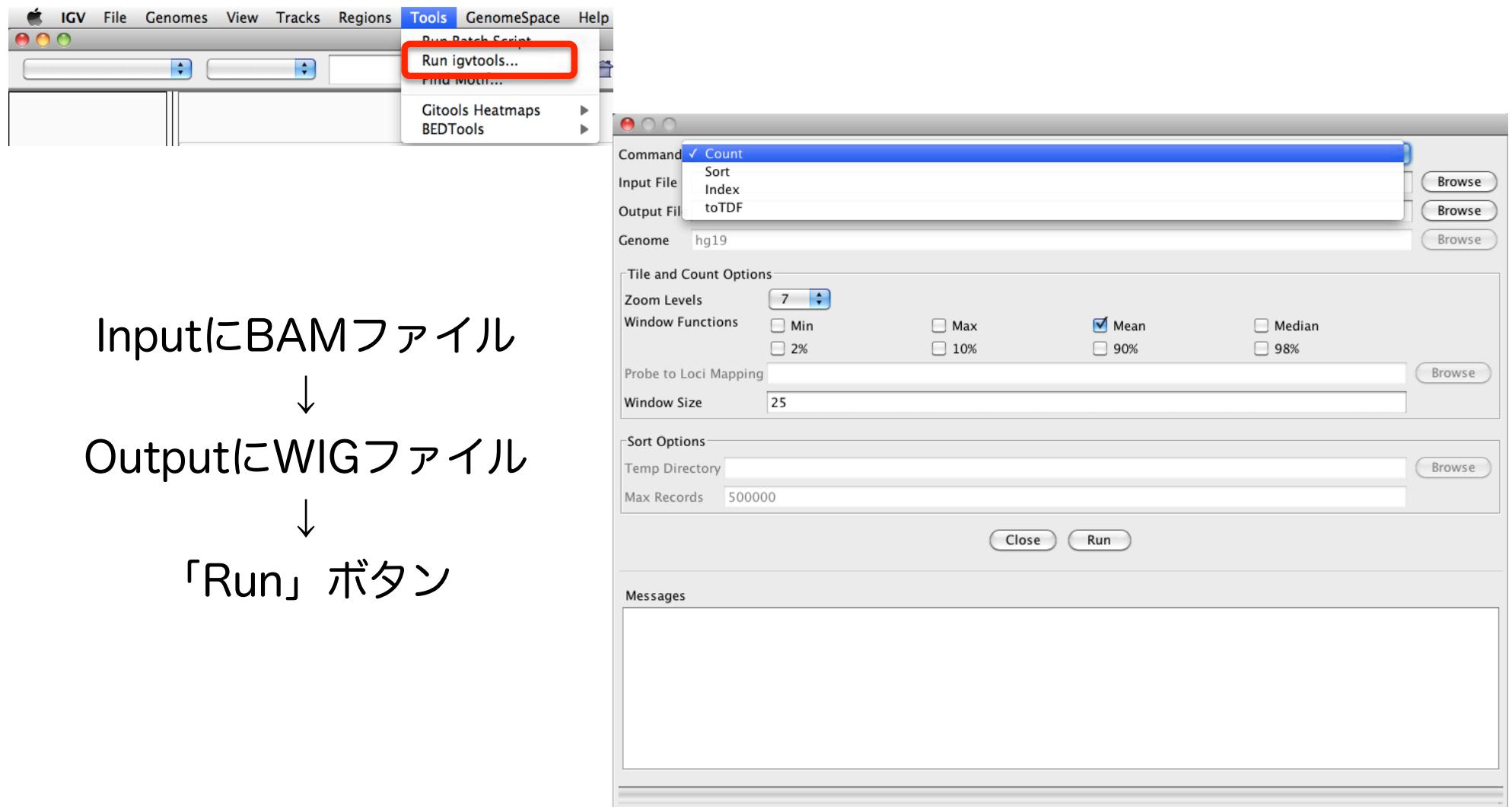
OutputにTDFファイル



「Run」ボタン

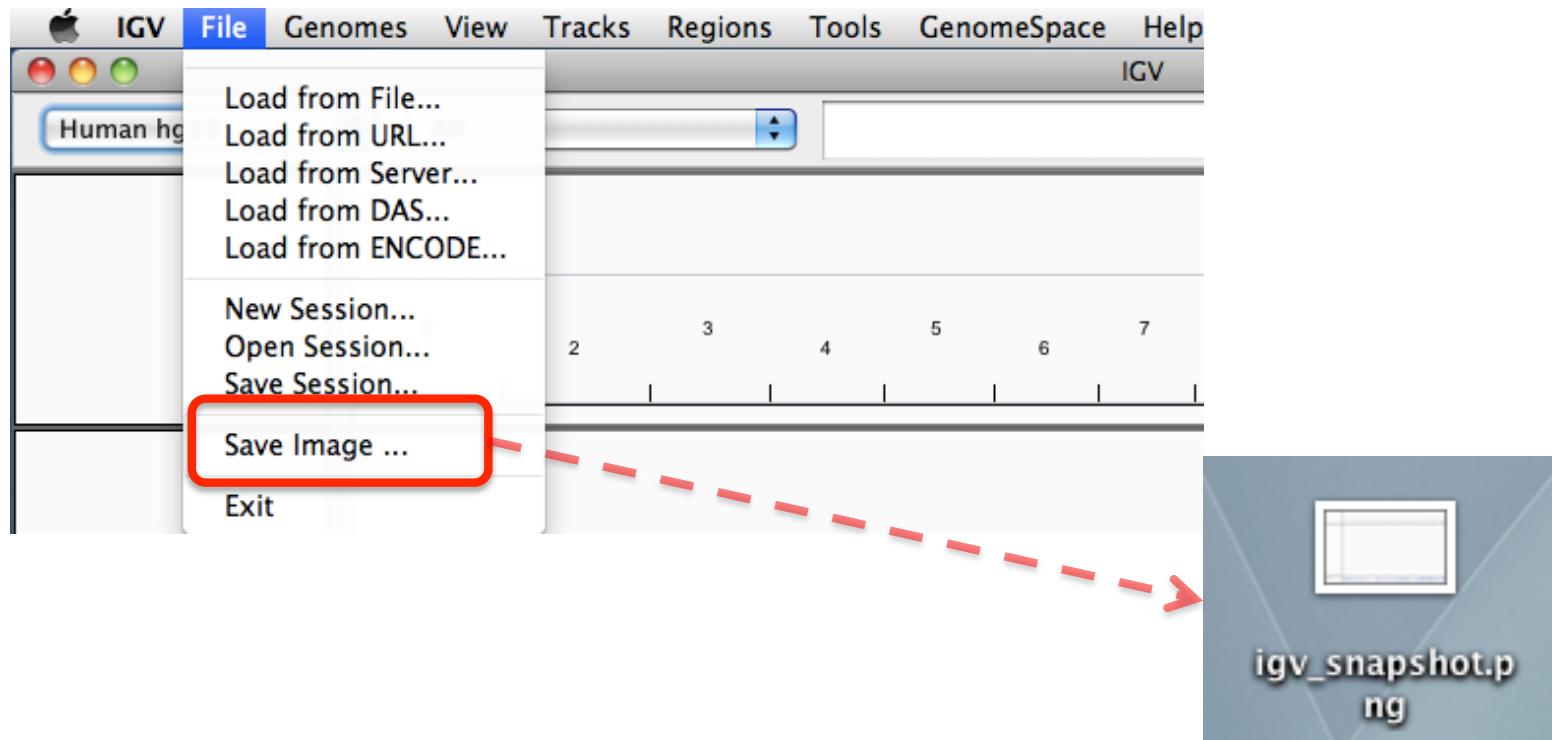
演習2：WIGファイルを作成

IGVToolsを利用してWIGファイルを作成してください。



演習3：表示画面を保存

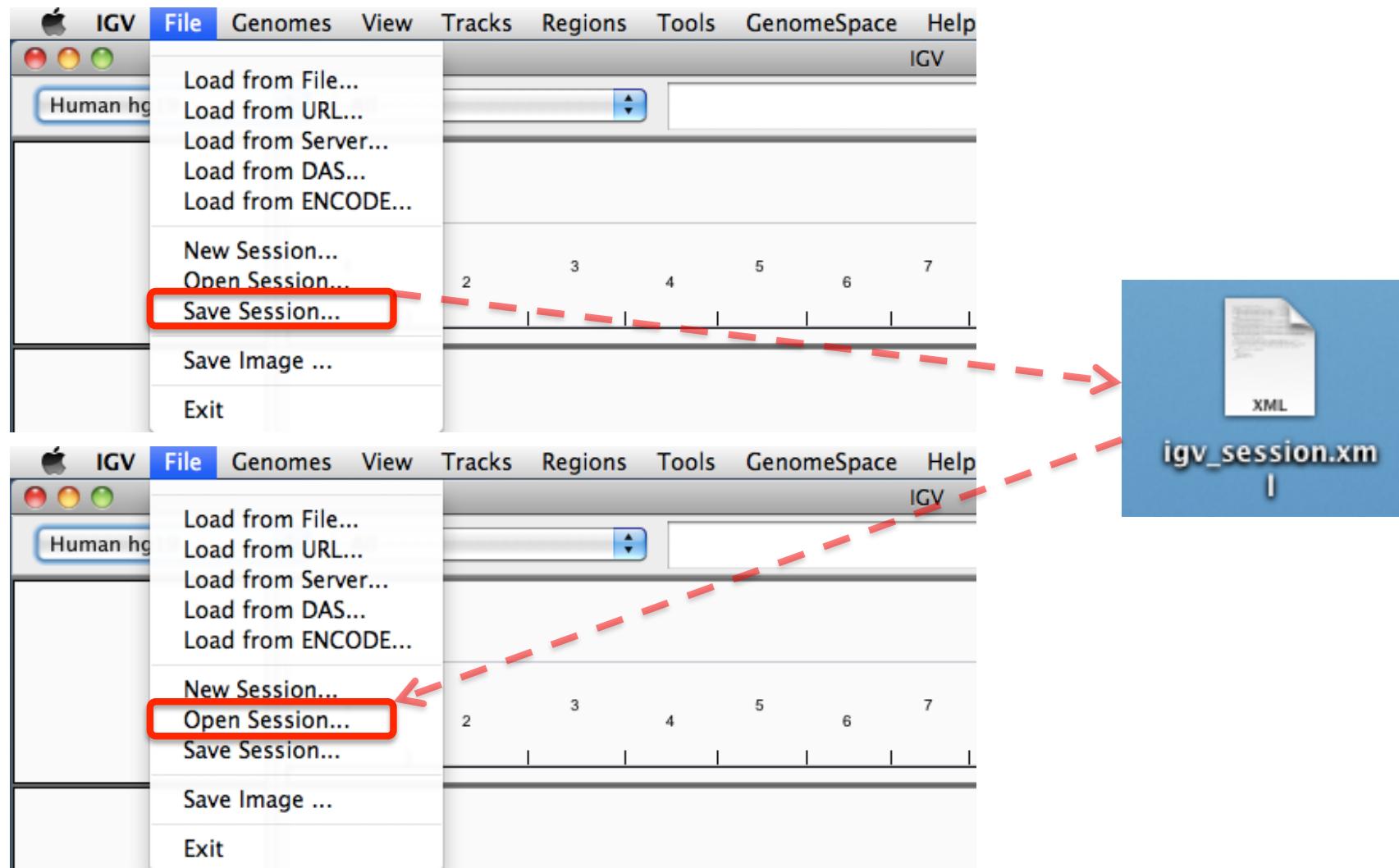
IGVで表示している画面を画像ファイル（PNGファイル）として保存してください。。



論文に掲載する図を作成することができます。

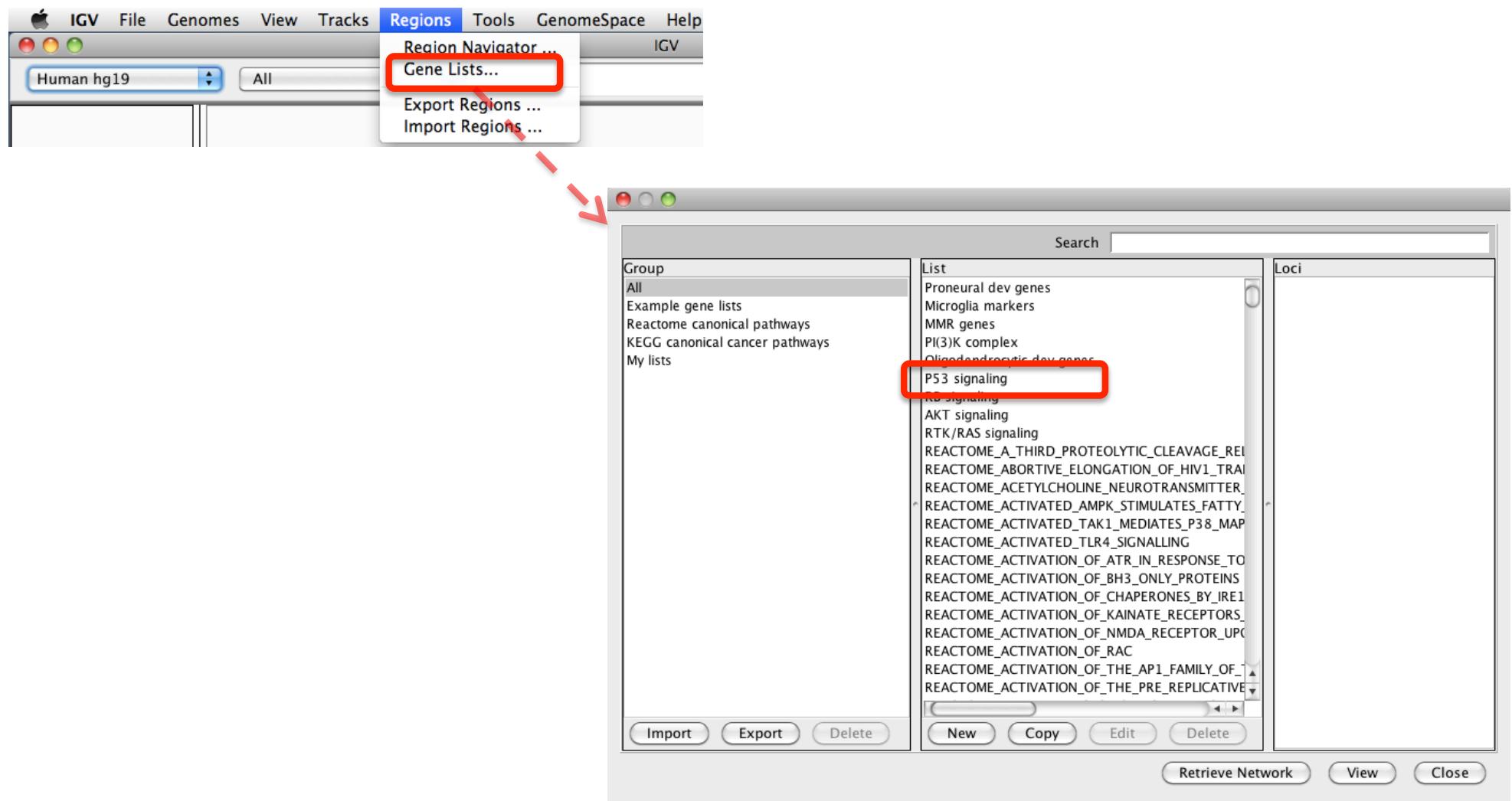
演習 4 : IGV sessionの保存と呼出

これまでIGVで読み込んだファイルや操作内容をxml形式ファイルに保存してください。その後、保存したsession内容を呼出してください。



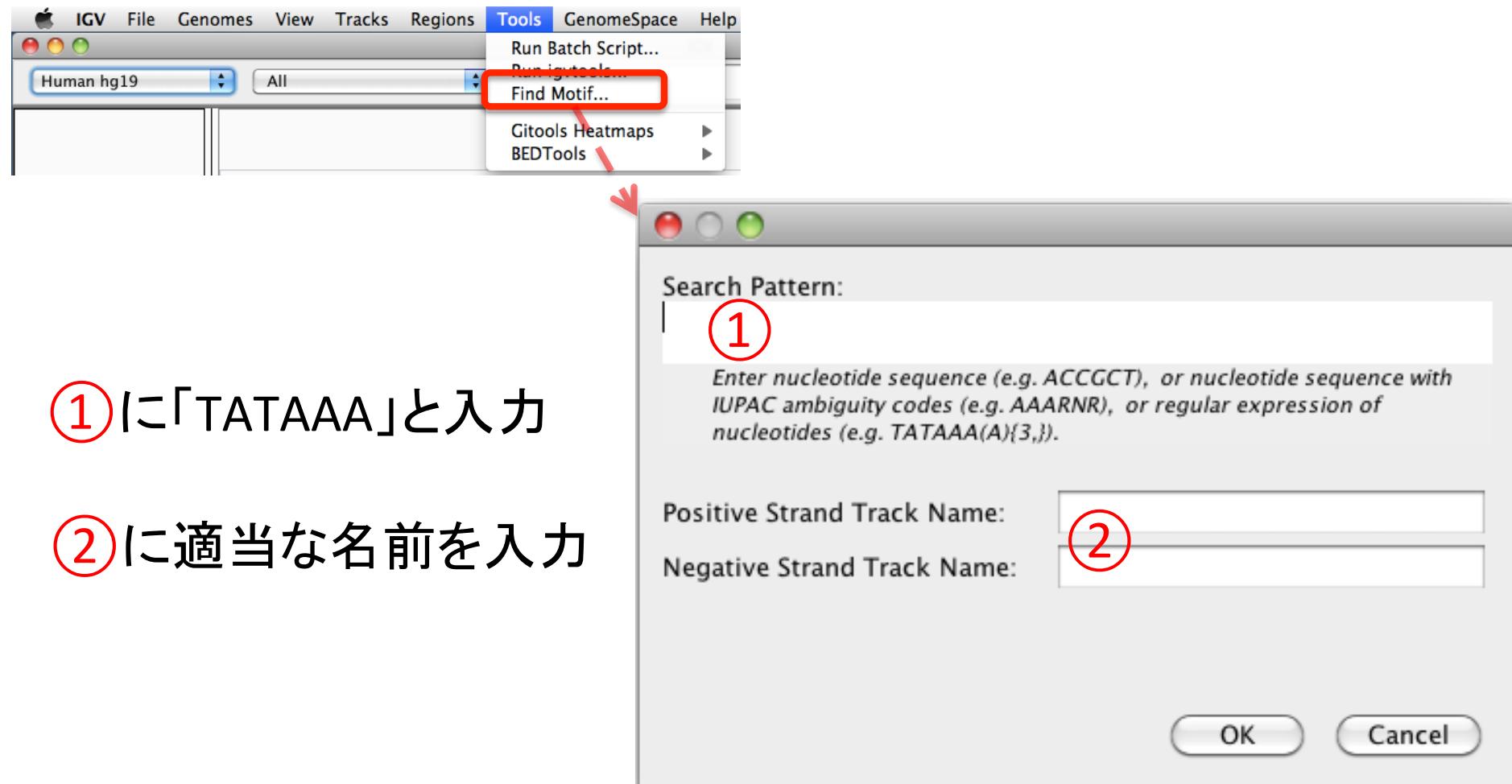
演習5：遺伝子セットの一覧表示

「Regions」タブの「Gene lists」を選択して、P53シグナルに関連する全ての遺伝子のローカスを確認してみましょう。



演習6：塩基モチーフ配列解析

「Tools」タブの「Find Motif」を選択して、転写因子のコンセンサス配列 'TATAAAA' と一致するゲノム領域を探してみましょう。



演習7：ENCODEデータの入手

ENCODEプロジェクトで解析されたデータを入手して、表示することができます。ファイル形式に注意しながら、データをダウンロードしましょう。なお、一部のゲノムバージョン（hg19, mm10）しか対応していません。

The image shows two screenshots of the IGV (Integrating Genome Viewer) software. The left screenshot shows the 'File' menu with the 'Load from ENCODE...' option highlighted by a red box. The right screenshot shows a 'Encode Production Data' dialog box with a 'Filter:' input field also highlighted by a red box. The dialog lists 19,954 rows of data, with columns including cell, datatype, antibody, view, replicate, type, lab, and hub. A red arrow points from the 'Load from ENCODE...' menu item to the 'Filter:' input field.

IGV

File Genomes View Tracks Regions Tools GenomeSpace Help

Human hg

Load from File...
Load from URL...
Load from Server...
Load from DAS
Load from ENCODE...
New Session...
Open Session...
Save Session...
Save Image ...
Exit

IGV

Encode Production Data

Filter: 19,954 rows

cell	datatype	antibody	view	replicate	type	lab	hub
89881	DnaseSeq	Peaks			narrowPeak	Duke	Data
AoSMC	DnaseSeq	Peaks			narrowPeak	Duke	Data
Chorion	DnaseSeq	Peaks			narrowPeak	Duke	Data
CLL	DnaseSeq	Peaks			narrowPeak	Duke	Data
Fibrobl	DnaseSeq	Peaks			narrowPeak	Duke	Data
FibroP	DnaseSeq	Peaks			narrowPeak	Duke	Data
Gliobla	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM12891	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM12892	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM18507	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM19238	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM19239	DnaseSeq	Peaks			narrowPeak	Duke	Data
GM19240	DnaseSeq	Peaks			narrowPeak	Duke	Data
H9ES	DnaseSeq	Peaks			narrowPeak	Duke	Data
HeLa-S3	DnaseSeq	Peaks			narrowPeak	Duke	Data
Hepatocyt...	DnaseSeq	Peaks			narrowPeak	Duke	Data
HPDE6-E...	DnaseSeq	Peaks			narrowPeak	Duke	Data
HSMM_emb	DnaseSeq	Peaks			narrowPeak	Duke	Data
HTR8svn	DnaseSeq	Peaks			narrowPeak		Data
Huh-7.5	DnaseSeq	Peaks			narrowPeak	Duke	Data
Huh-7	DnaseSeq	Peaks			narrowPeak	Duke	Data
iPS	DnaseSeq	Peaks			narrowPeak	Duke	Data
Ishikawa	DnaseSeq	Peaks			narrowPeak	Duke	Data
Ishikawa	DnaseSeq	Peaks			narrowPeak	Duke	Data
LNCaP	DnaseSeq	Peaks			narrowPeak	Duke	Data
MCF-7	DnaseSeq	Peaks			narrowPeak	Duke	Data
Medullo	DnaseSeq	Peaks			narrowPeak	Duke	Data
Melano	DnaseSeq	Peaks			narrowPeak	Duke	Data
Myometr	DnaseSeq	Peaks			narrowPeak	Duke	Data
Osteobl	DnaseSeq	Peaks			narrowPeak	Duke	Data

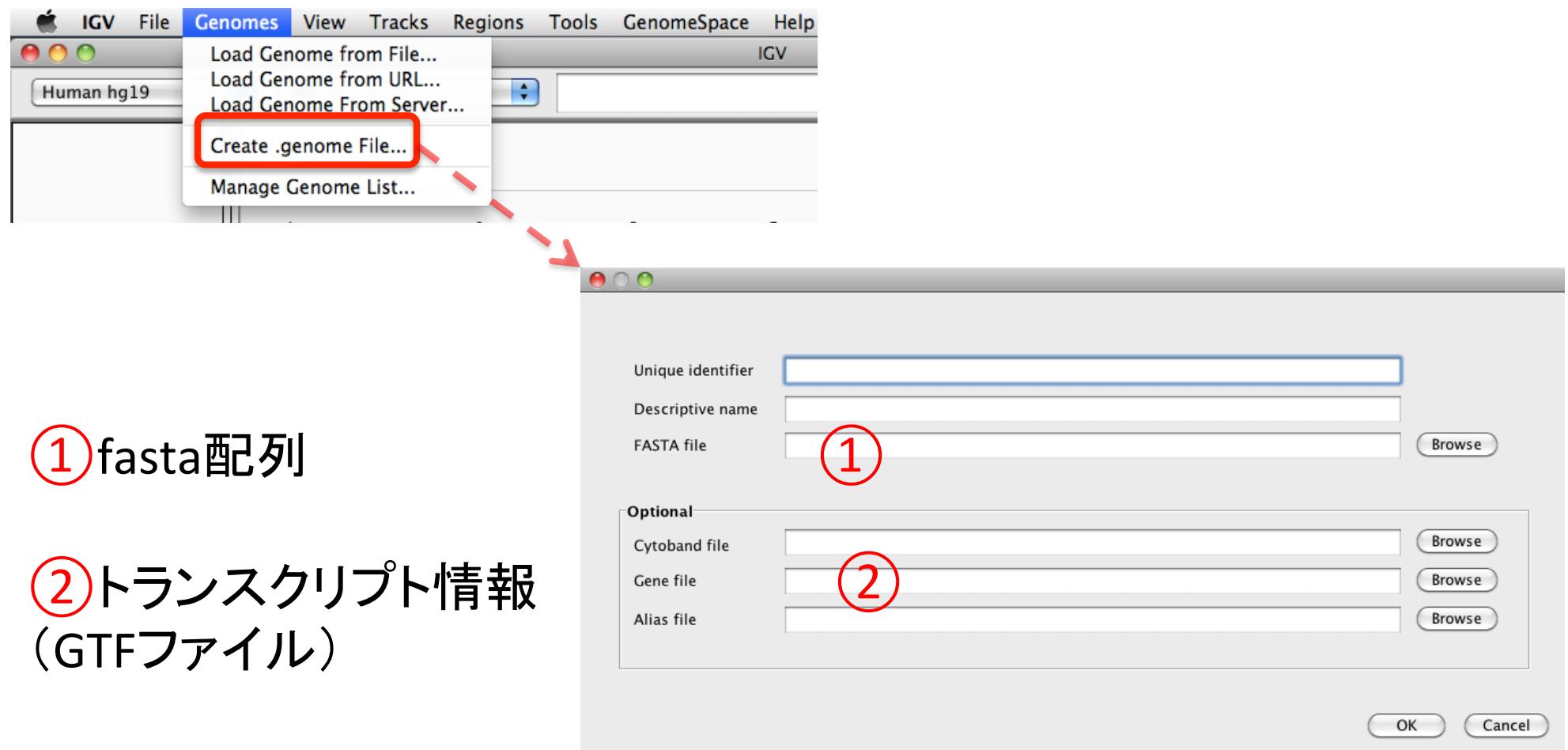
Load Cancel

キーワードでデータを
フィルタリング

例えば「ChIPseq
HeLaS3 bigWig」

補足

IGVでは用意されていない配列（新たなゲノムプロジェクトで決定されたゲノム配列）についても、IGV専用のgenomeファイルを作成することで、視覚化できます。



おわりに

視覚化ツールを利用して、

- 公共データベースに登録されているマッピングデータをご自身の研究にご活用ください。
- バイオインフォマティシャンにおまかせしている解析結果を見直しましょう。もし、視覚化データと結果が一致しないようであれば、バイオインフォマティシャンに問い合わせましょう。