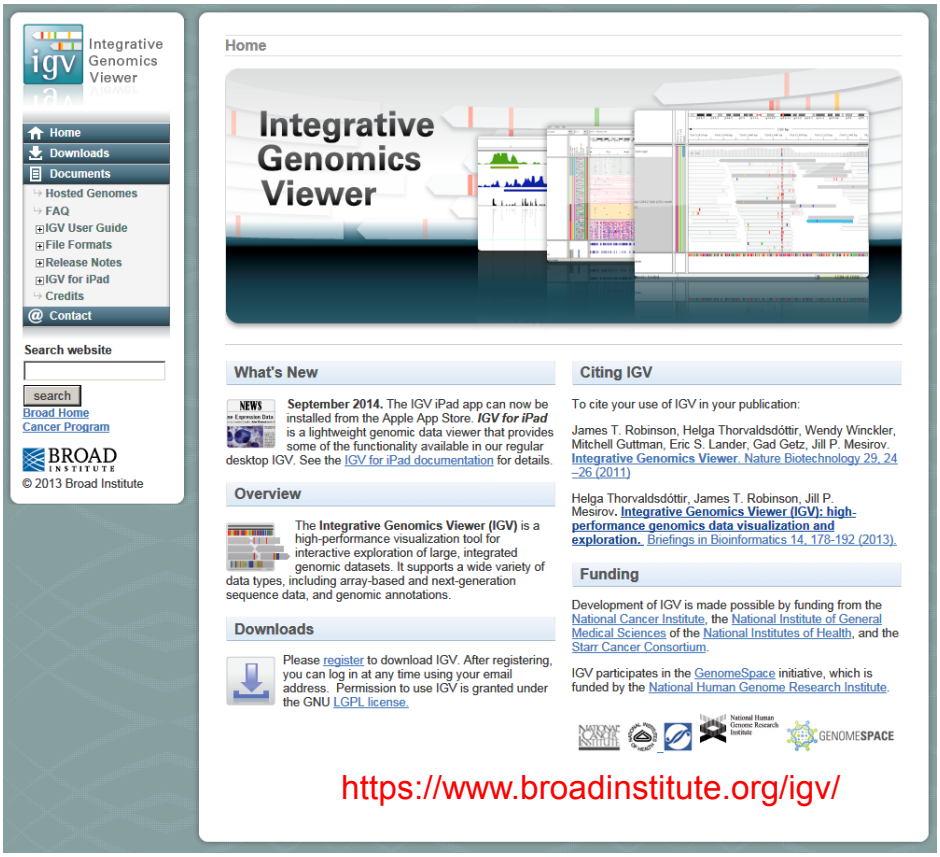
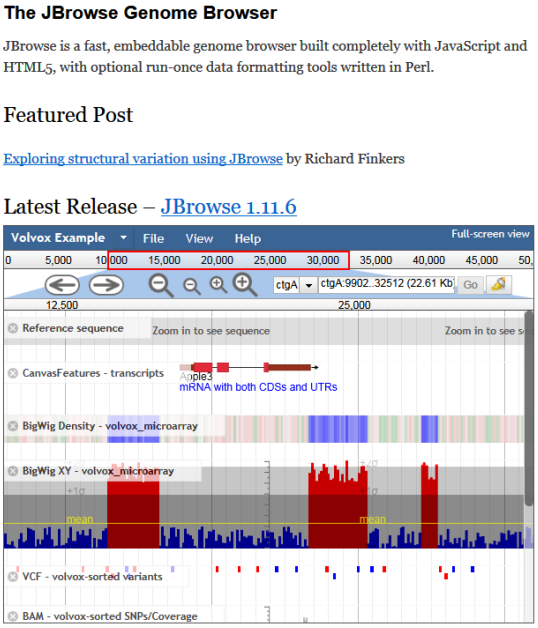
データ可視化ツール・IGVの紹介・実習



なぜIGVを取り上げるか

データ可視化ツール
・自分のパソコン（ローカル環境）にインストールして使うタイプ
・サーバーに構築して、ネットワークで使うタイプ



コミュニティーに広く利用、あるいは
ウェブ公開を目的とするには良いが、
ネットワーク・情報セキュリティの
高度な知識も要求される。

より大容量なデータに対応できる。

管理者的な人がいて、その人がやって
くれるなら、これも良いが。

もっとお手軽なものとしてIGVを紹介

可視化ツールに求められるものは何か

膨大なデータを如何に直感的に理解できるようにするか
sortや絞り込みができる表データと対比双璧

・配列、GC ratio、遺伝子情報
・遺伝子発現情報
・SNPの位置情報・頻度情報
・様々なデータの精度情報

レファレンス配列 / gene model / gene annotationとNGSデータを並べて比較
複数のデータセットを並べて比較

色々なデータ（variant, 発現, ChIP, BSseq等々）を、様々なスケールで
比較・統合的に解釈できるようにしたい

ゲノムviewerに自分のデータを乗せ、
統合的直感的に解釈できること

可視化ツールをどう選ぶか

選択の基準
genome data viewing に求められるもの
取捨選択の基準

1. 無料 / 有料 / 基本無料

2. 個人的レベルの使用 / コミュニティーレベルの使用

3. 見るだけ/自分から色々工夫

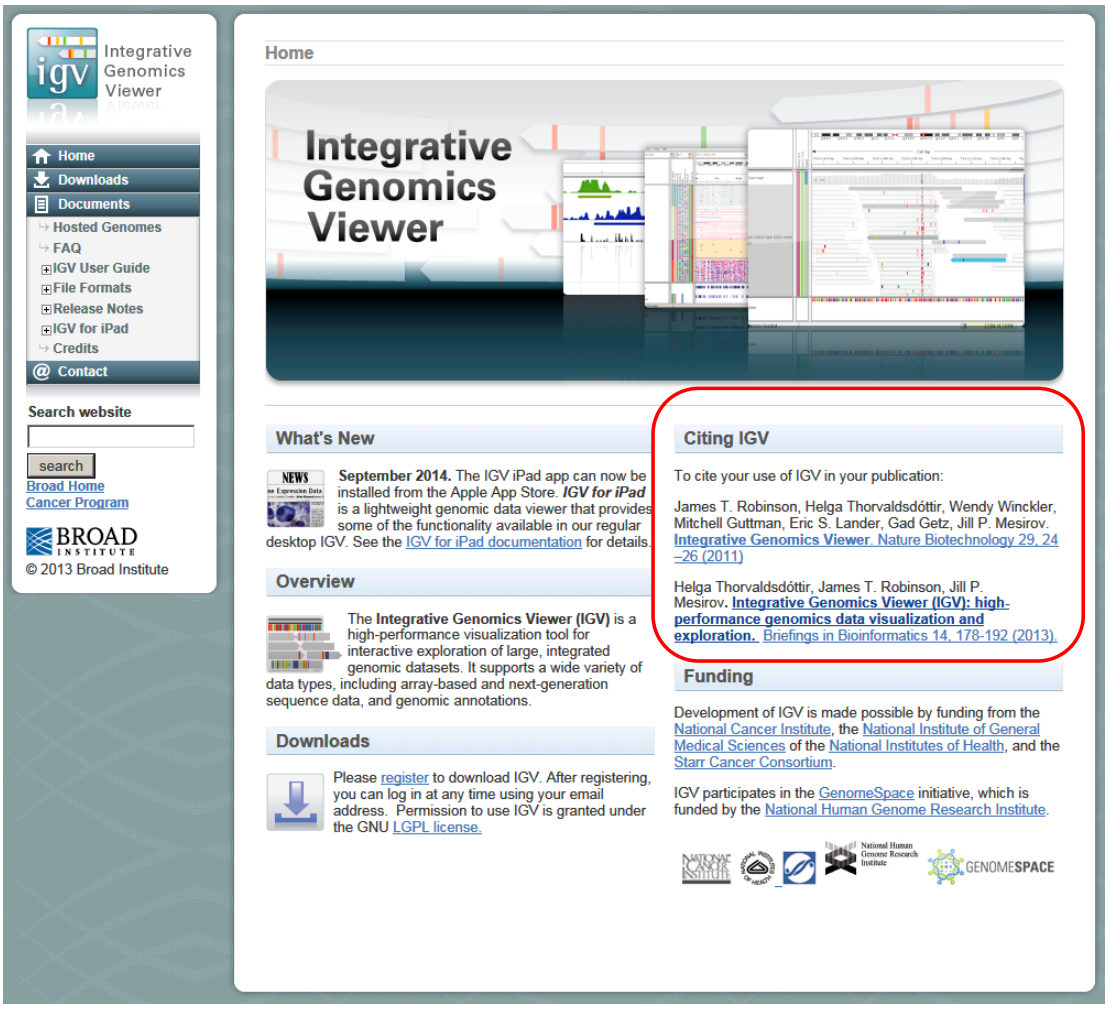
4.アクセスのしやすさ・使いやすさ
導入に必要なコンピュータスペック
マニュアルは分かりやすいか
情報の多さ
利用の簡便さ
使っている人が近くにいるか

Integrative Genomics Viewer(IGV)

お手軽ツール

・アカデミックウェアで無料
・コミュニティーでの利用者が多いから、情報も多い
・javaのプログラムなので、オールプラットフォーム対応
・マニュアルは親切、サンプルデータのある
・WEBサーバーではなく、PCレベルでできる
・データ閲覧環境の共有が可能

誰もが簡便に使えるものが良い。



nature biotechnology

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Integrative genomics viewer

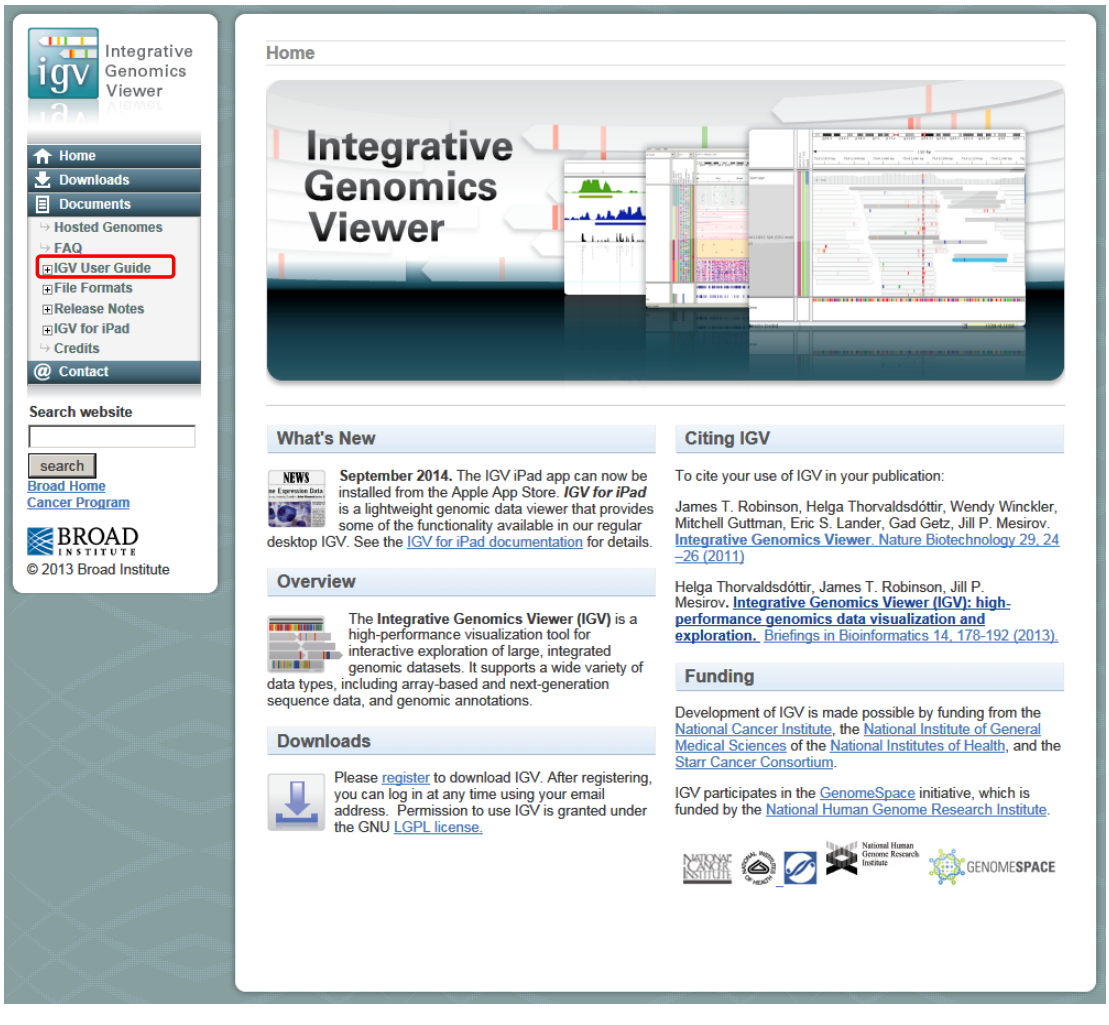
James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz & Jill P Mesirov

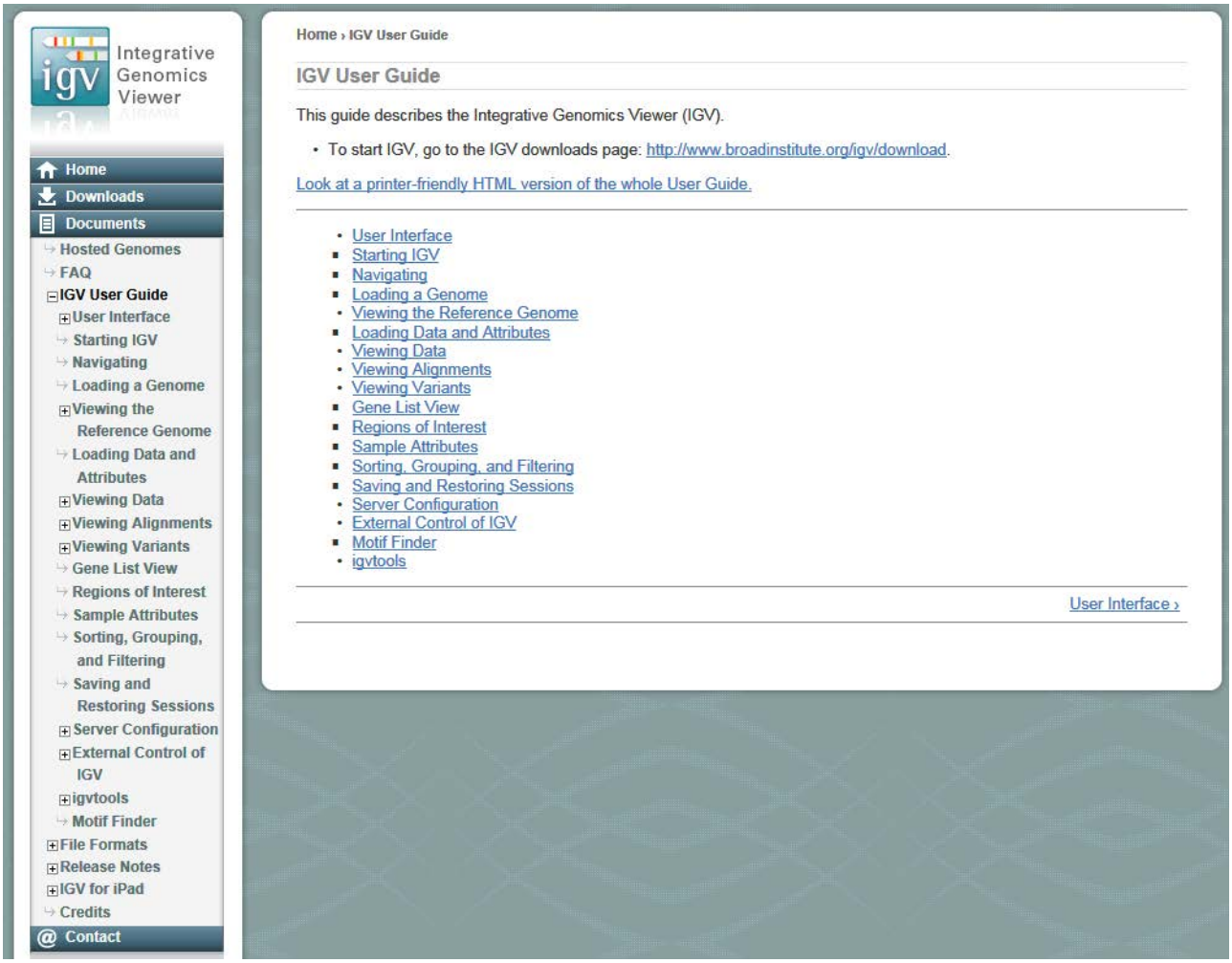
Affiliations | Corresponding authors

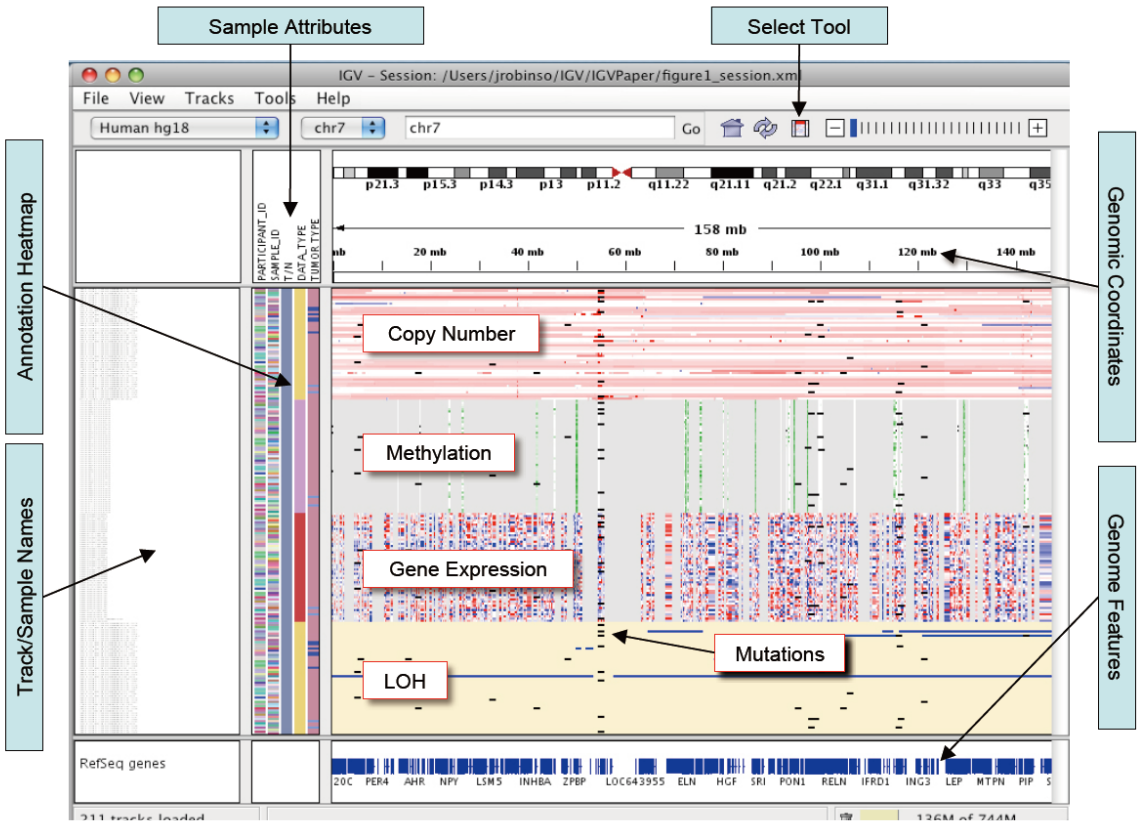
Nature Biotechnology **29**, 24–26 (2011) | doi:10.1038/nbt.1754
Published online 10 January 2011

To the Editor:

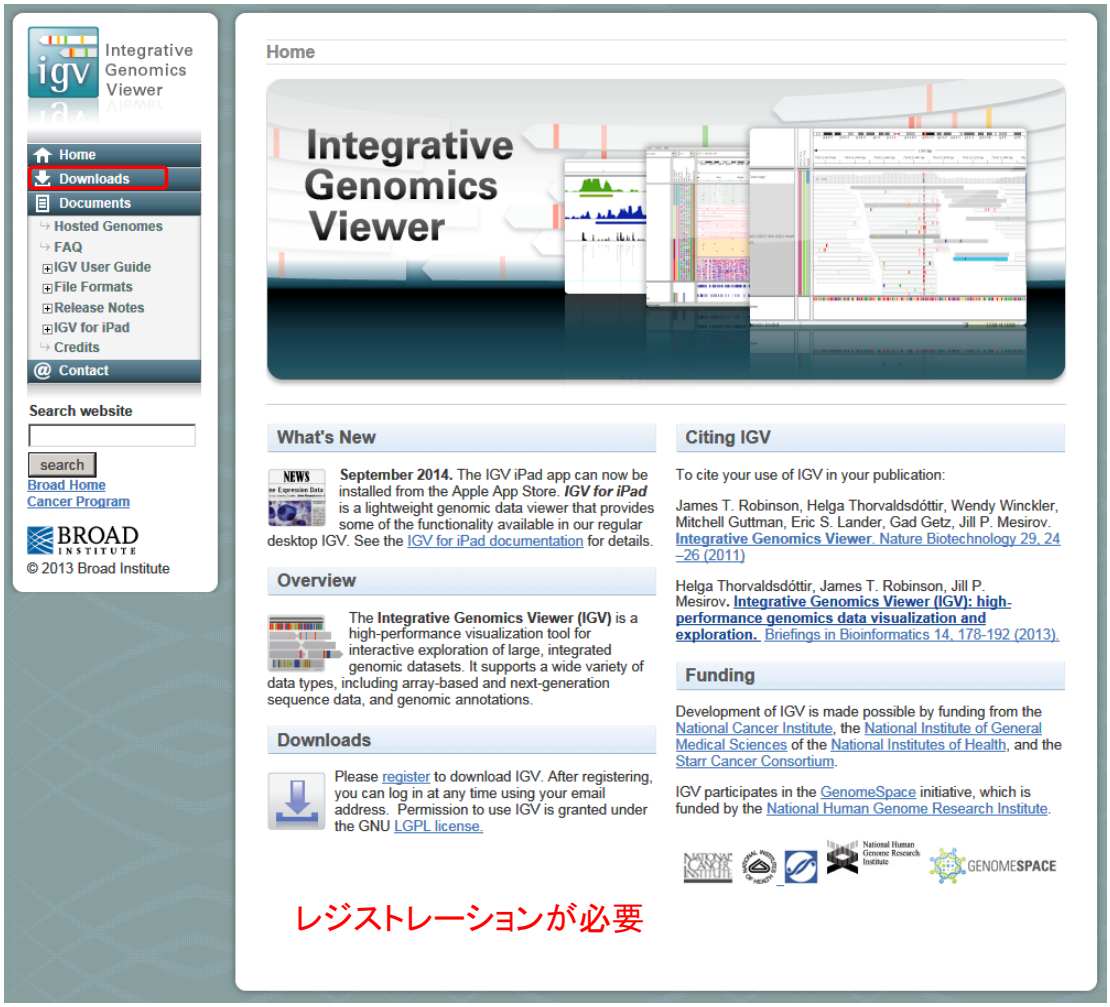
Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Analysis of these large, diverse data sets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large data sets and to flexibly integrate multiple data types, including clinical data. However, the sheer volume and scope of data pose a significant challenge to the development of such tools.

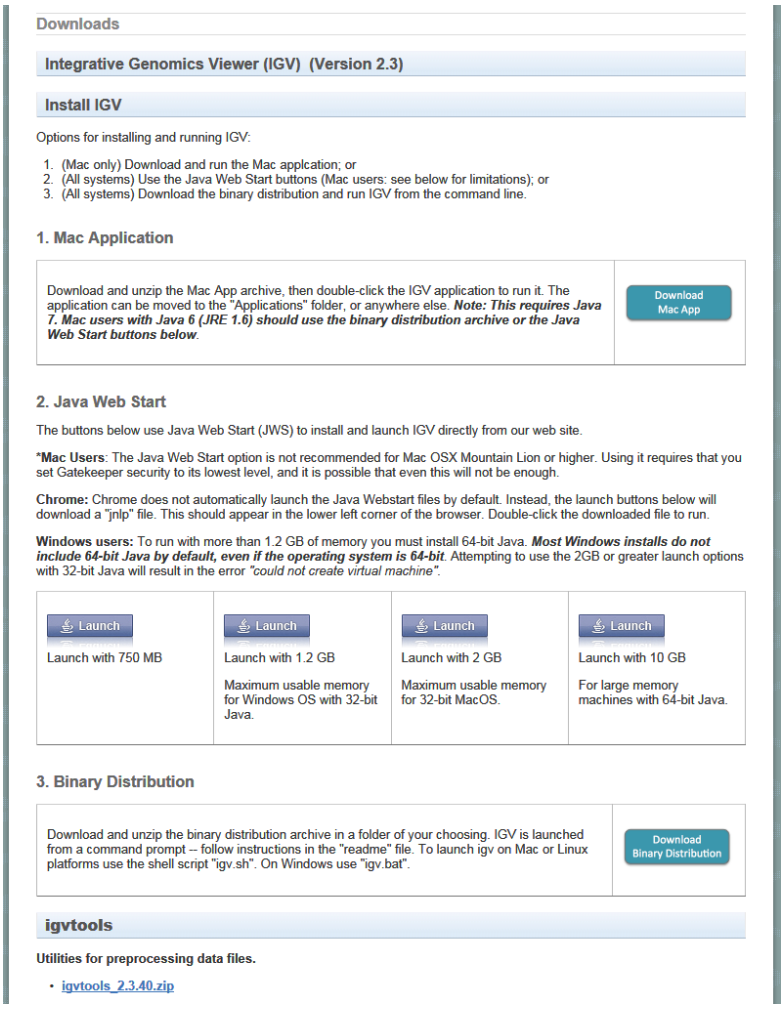


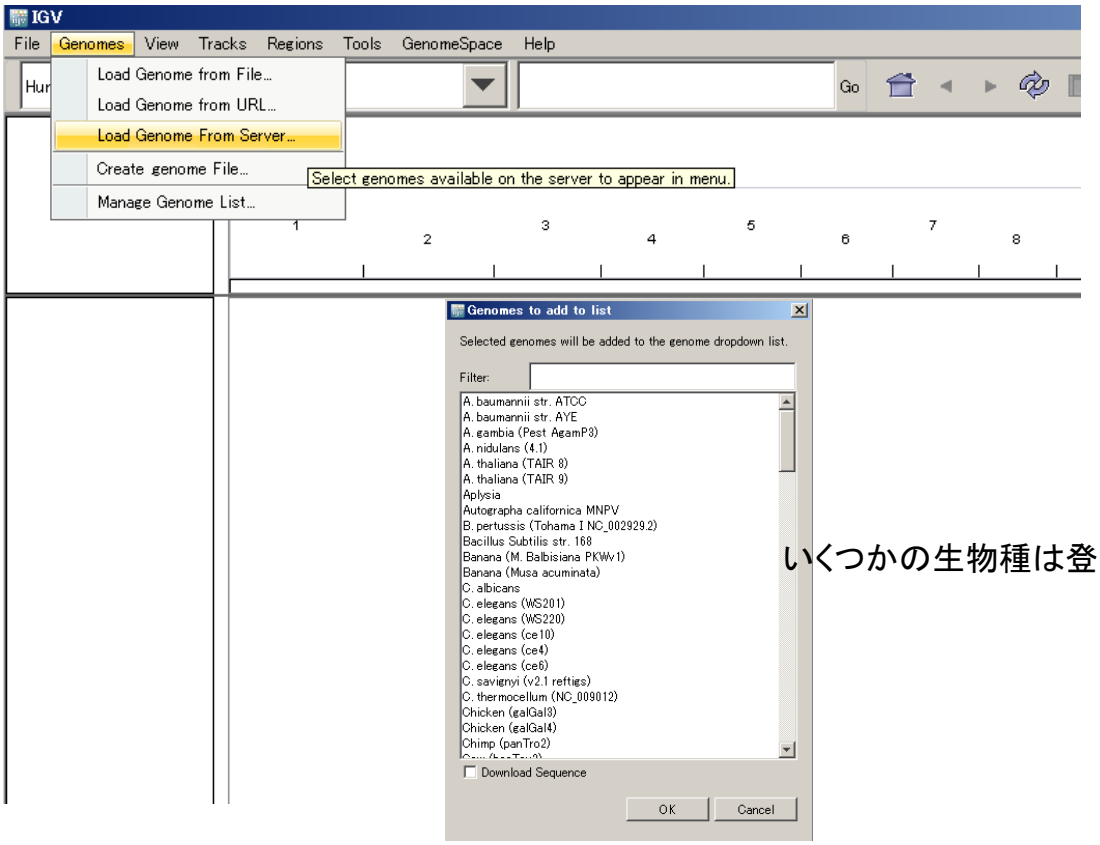




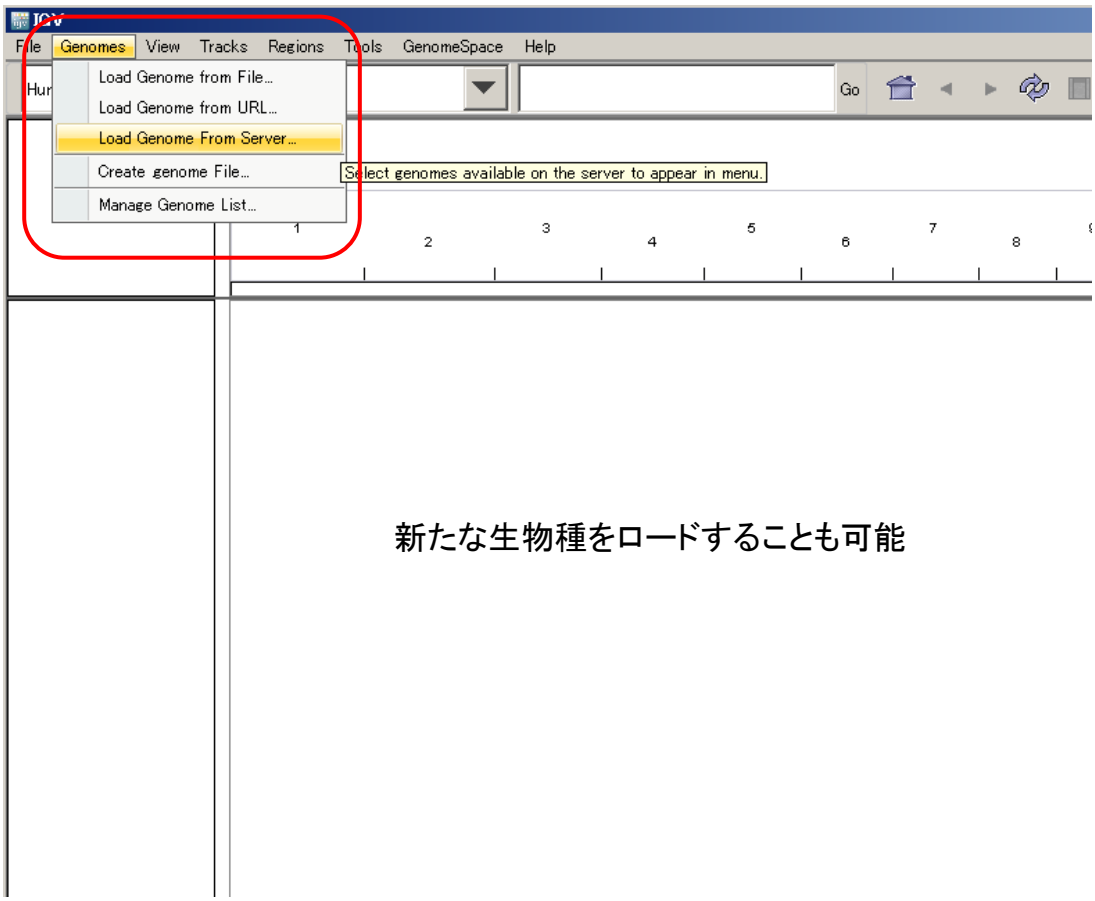
Nature Biotech. 29:24–26 (2011) Supplement figureからの抜粋


Integrative Genomics Viewer
Home
Downloads
Documents
Hosted Genomes
FAQ
IGV User Guide
File Formats
Release Notes
IGV for iPad
Credits
Contact
Search website
search
Broad Home
Cancer Program
BROAD INSTITUTE
© 2013 Broad Institute
Home
Integrative Genomics Viewer
What's New
September 2014. The IGV iPad app can now be installed from the Apple App Store. IGV for iPad is a lightweight genomic data viewer that provides some of the functionality available in our regular desktop IGV. See the IGV for iPad documentation for details.
Overview
The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.
Downloads
Please register to download IGV. After registering, you can log in at any time using your email address. Permission to use IGV is granted under the GNU LGPL license.
Citing IGV
To cite your use of IGV in your publication:
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)
Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).
Funding
Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.
IGV participates in the GenomeSpace initiative, which is funded by the National Human Genome Research Institute.
GENOMESPACE
レジストレーションが必要


Downloads
Integrative Genomics Viewer (IGV) (Version 2.3)
Install IGV
Options for installing and running IGV:
1. (Mac only) Download and run the Mac application; or
2. (All systems) Use the Java Web Start buttons (Mac users: see below for limitations); or
3. (All systems) Download the binary distribution and run IGV from the command line.
1. Mac Application
Download and unzip the Mac App archive, then double-click the IGV application to run it. The application can be moved to the "Applications" folder, or anywhere else. Note: This requires Java 7. Mac users with Java 6 (JRE 1.6) should use the binary distribution archive or the Java Web Start buttons below.
Download Mac App
2. Java Web Start
The buttons below use Java Web Start (JWS) to install and launch IGV directly from our web site.
*Mac Users: The Java Web Start option is not recommended for Mac OSX Mountain Lion or higher. Using it requires that you set Gatekeeper security to its lowest level, and it is possible that even this will not be enough.
Chrome: Chrome does not automatically launch the Java Webstart files by default. Instead, the launch buttons below will download a "jnlp" file. This should appear in the lower left corner of the browser. Double-click the downloaded file to run.
Windows users: To run with more than 1.2 GB of memory you must install 64-bit Java. Most Windows installs do not include 64-bit Java by default, even if the operating system is 64-bit. Attempting to use the 2GB or greater launch options with 32-bit Java will result in the error "could not create virtual machine".
Launch
Launch with 750 MB
Launch
Launch with 1.2 GB
Maximum usable memory for Windows OS with 32-bit Java.
Launch
Launch with 2 GB
Maximum usable memory for 32-bit MacOS.
Launch
Launch with 10 GB
For large memory machines with 64-bit Java.
3. Binary Distribution
Download and unzip the binary distribution archive in a folder of your choosing. IGV is launched from a command prompt -- follow instructions in the "readme" file. To launch igv on Mac or Linux platforms use the shell script "igv.sh". On Windows use "igv.bat".
Download Binary Distribution
igvtools
Utilities for preprocessing data files.
igvtools_2.3.40.zip



いくつかの生物種は登録されている



新たな生物種をロードすることも可能

ゲノムViewerなので次世代DNAシーケンサーのデータに限定されない。
マイクロアレイの結果や、ゲノムアノテーションの情報も随時表示できる。

対応するファイル形式に応じて、表示方法が決まる。

File Formats

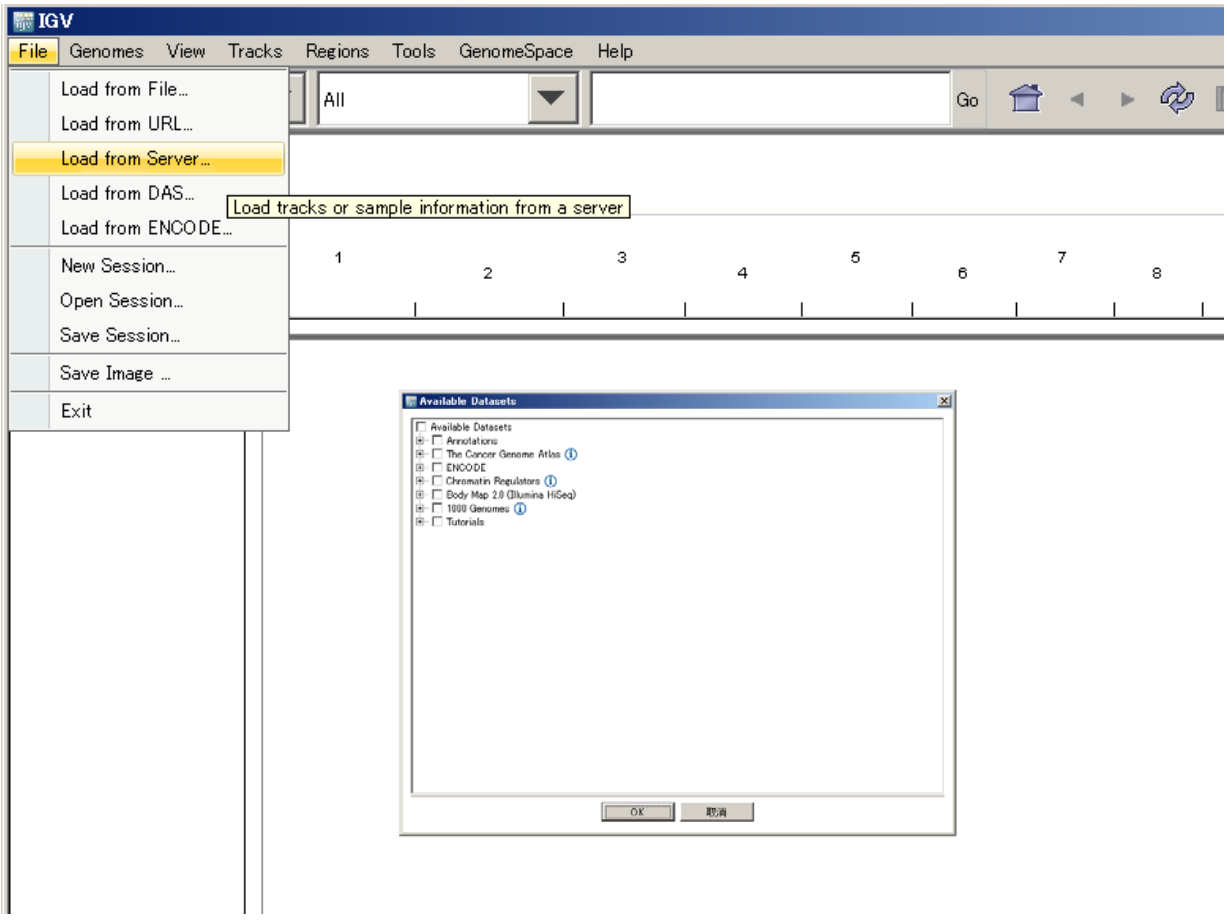
- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG
- chrom.sizes

File Formats

IGV supports a number of different file formats for experimental data and genome annotations. For a complete list of supported formats see http://www.broadinstitute.org/igv/FileFormats. The following table shows the recommended file formats for a number of common data types.

Source Data	Recommended File Formats
ChIP-Seq, RNA-Seq	WIG, TDF
Copy number	CN, SNP, TDF, canary_calls (Birdsuite)
Gene expression data	GCT, RES, TDF
Genome annotations	GFF, BED, GTF, PSL, UCSC table format
GISTIC data	GISTIC
LOH data	LOH, TDF
Mutation data	MUT, MAF
Variant calls	VCF
RNAi data	GCT
Segmented data	SEG, CBS
Sequence alignment data	BAM, SAM, PSL
Any numeric data	IGV, WIG, TDF
Sample metatadata	Tab-delimited sample info file

公開情報のviewerとして



その他の便利機能

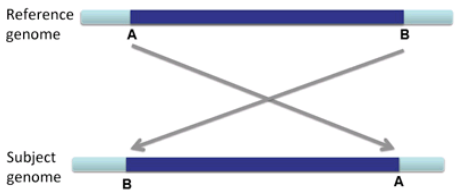
セッションの保存
　表示しているデータの読み込み状況を、それごと保存。
　セッションをロードすることで、意図した画面を表示できる。
　　データセットが揃っていること、フォルダー構造が同一である必要がある。

バッチ処理
　重要領域の画面スナップショットを自動で取ったりできる。

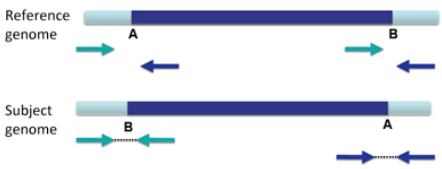
```
new
load myfile.bam
snapshotDirectory mySnapshotDirectory
genome hg18
goto chr1:65,289,335-65,309,335
sort position
collapse
snapshot
goto chr1:113,144,120-113,164,120
sort base
collapse
snapshot
```

Inversions

An inversion is a large section of DNA that is reversed in the subject genome compared to the reference genome.



When an inversion shows up in paired-end reads, the reads are distinctively variant from the reference genome.



This appears in IGV as shown below.

Interpreting Color by Insert Size

The inferred insert size can be used to detect structural variants, such as:

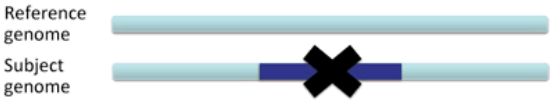
- deletions
- insertions
- inter-chromosomal rearrangements

IGV uses color coding to flag anomalous insert sizes. When you select Color alignments>by insert size in the popup menu, the default coloring scheme is:

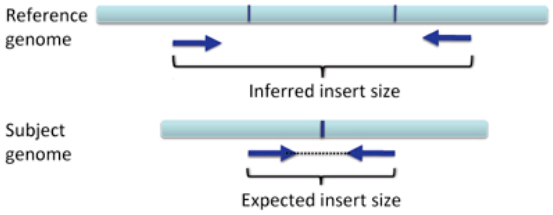
- for an insert that is larger than expected
- for an insert that is smaller than expected
- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y for paired end reads that are coded by the chromosome on which their mates can be found

Deletions

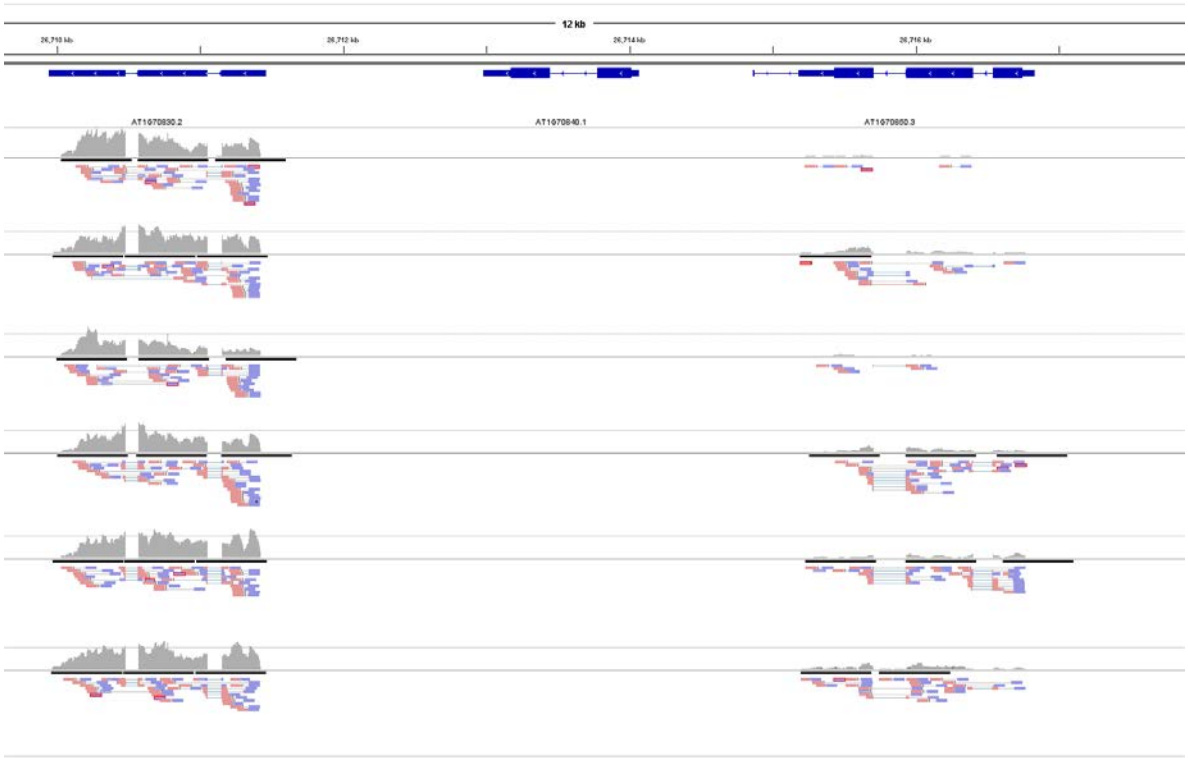
A deletion is a large section of DNA that is absent in the subject genome compared to the reference genome.



The "expected" insert size is the insert size obtained in sequencing the subject genome. The "inferred" insert size is the insert size that would result in the reference genome, assuming the same pair of reads.

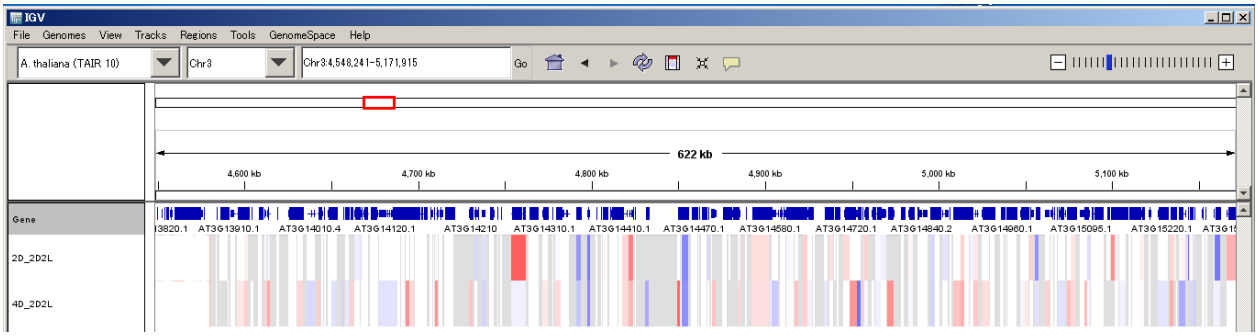


RNA-Seqのデータ表示させる

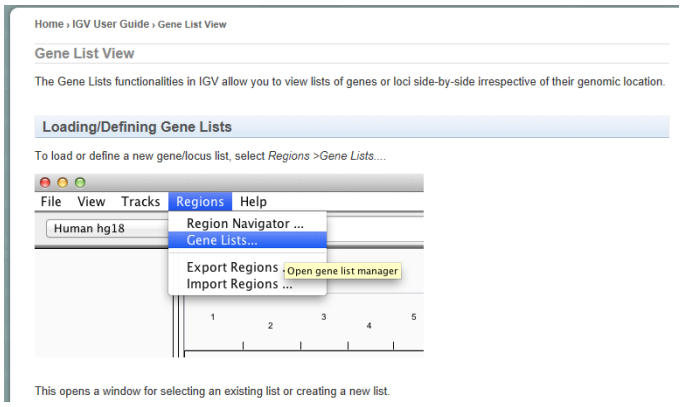


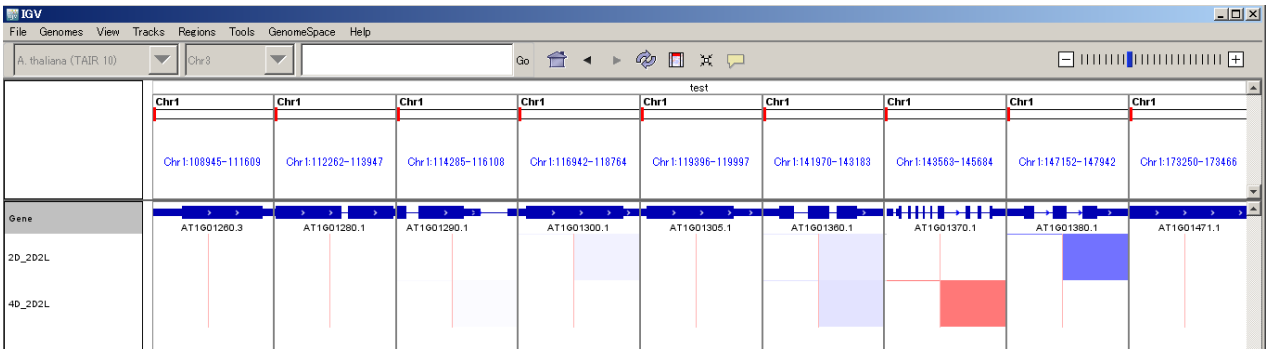
GCTファイルでgene ローカスの発現情報を図示

```
#
#
Name        Description              2D_2D2L      4D_2D2L
ANAC001     |@Chr1:3630-5899         -2.60184     -2.60956
DCL1        |@Chr1:23145-33153       -0.742675    -1.5642
MIR838A     |@Chr1:23145-33153       0            0
AT1G01073   |@Chr1:44676-44787       0            0
IQD18       |@Chr1:52238-54692       -1.93871     -1.13128
AT1G01115   |@Chr1:56623-56740       0            0
GIF2        |@Chr1:72338-74737       -0.251287    -0.616679
AT1G01180   |@Chr1:75582-76758       0.45929      -0.809567
AT1G01210   |@Chr1:88897-89745       1.6964       0.857196
FKGP        |@Chr1:91375-95651       -0.174589    0.725947
AT1G01240   |@Chr1:99893-101834      -0.226384    -0.936641
AT1G01260   |@Chr1:108945-111609     -0.161848    0.315699
CYP703A2    |@Chr1:112262-113947     0            0
CNX3        |@Chr1:114285-116108     0.111249     -0.551359
AT1G01300   |@Chr1:116942-118764     -0.68348     0.108578
```

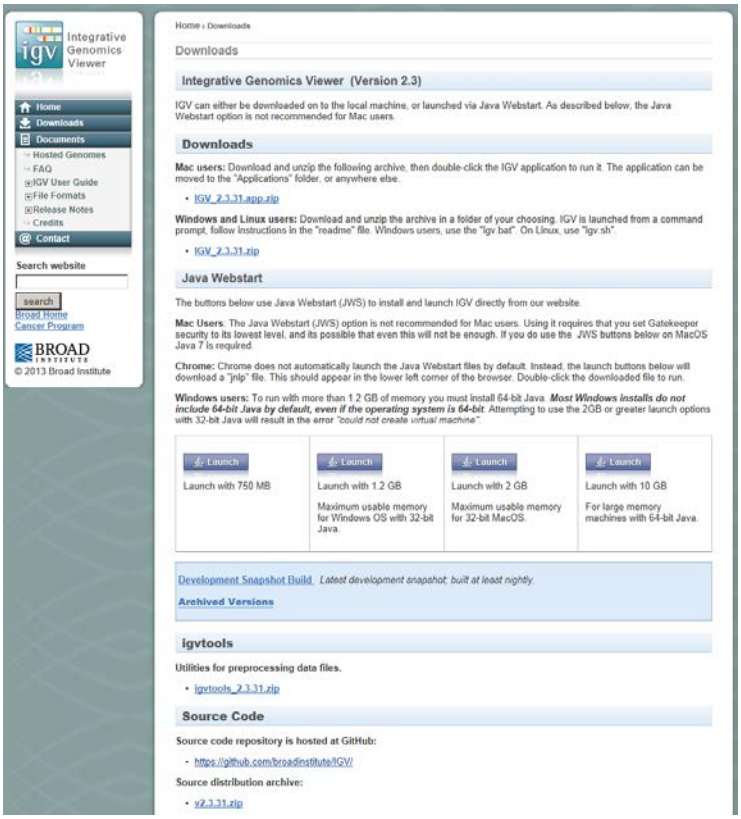


Gene listを定義して
サンプルごと
条件ごと
の発現・発現変動を
カラーマップできる





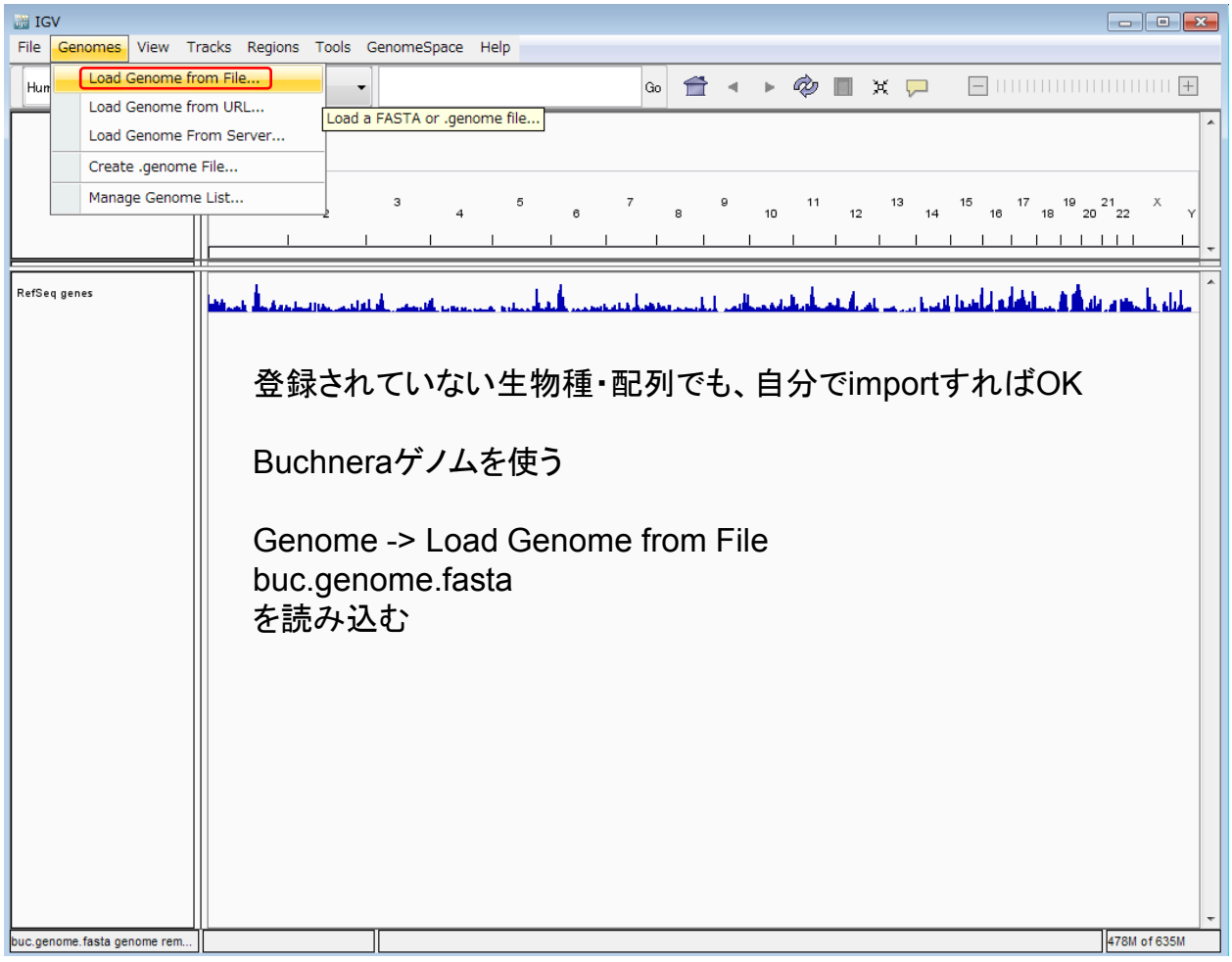
IGV実習



IGVの使用法を学ぶと共に
先のファイルフォーマットも
確認しよう

以下のファイルを確認

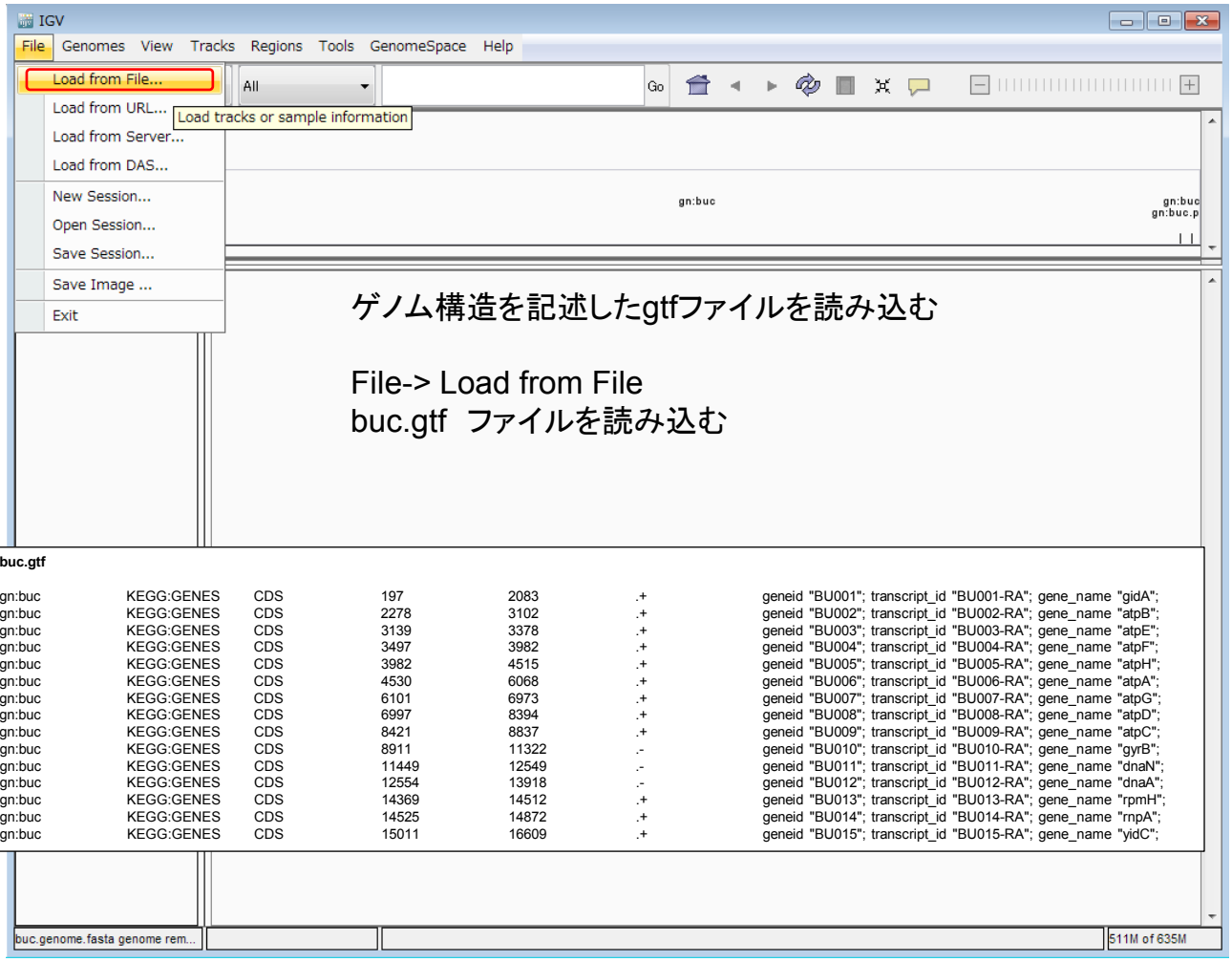
buc.genome.fasta
buc.gtf
buc_cg.wig
illumina_ex_B2_Read_bowtie2.mate.sort.bam
illumina_ex_B2_Read_bowtie2.mate.sort.bam.bai
illumina_ex_B4_Read_bowtie2.mate.sort.bam
illumina_ex_B4_Read_bowtie2.mate.sort.bam.bai



登録されていない生物種・配列でも、自分でimportすればOK

Buchneraゲノムを使う

Genome -> Load Genome from File
buc.genome.fasta
を読み込む

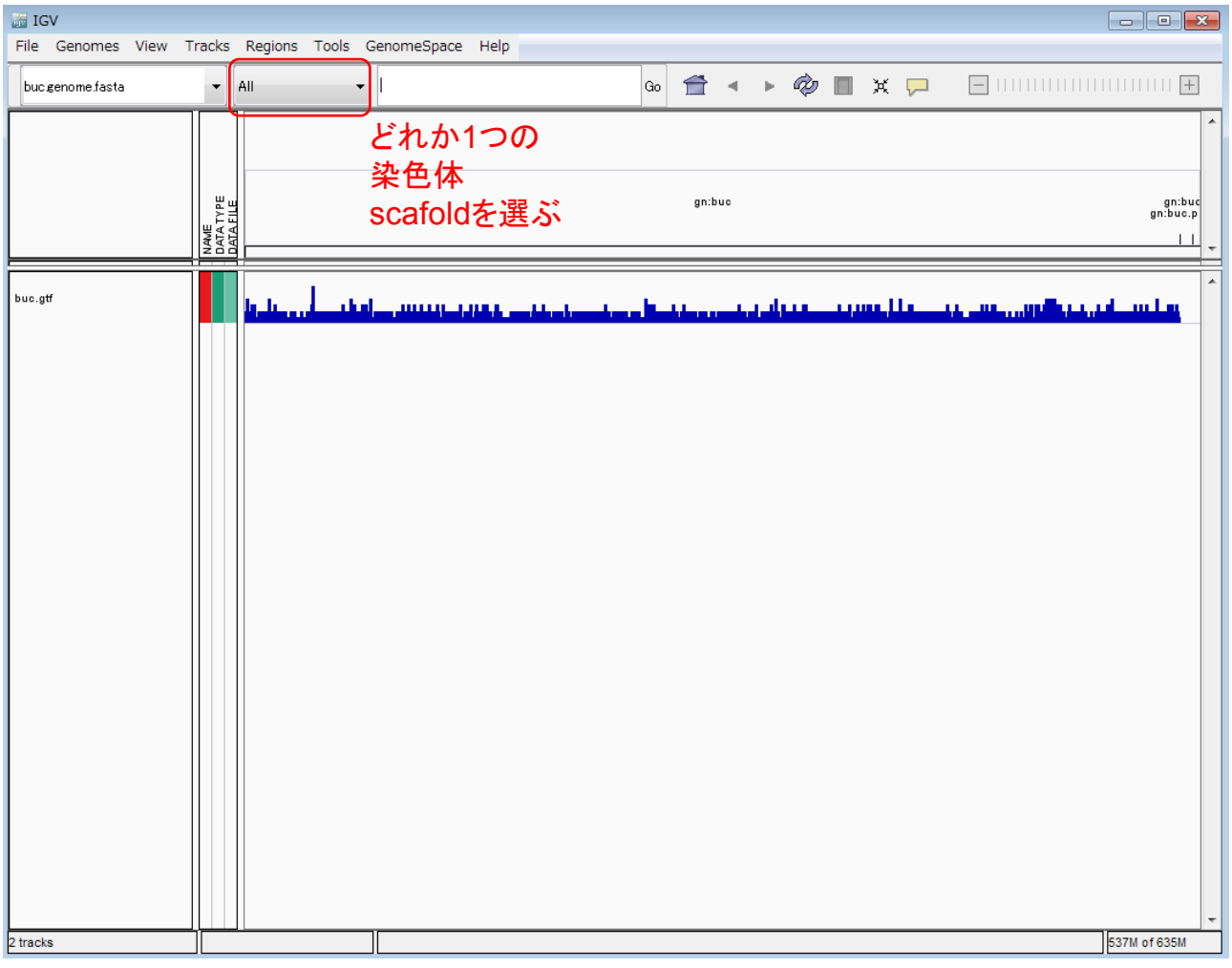


ゲノム構造を記述したgtfファイルを読み込む

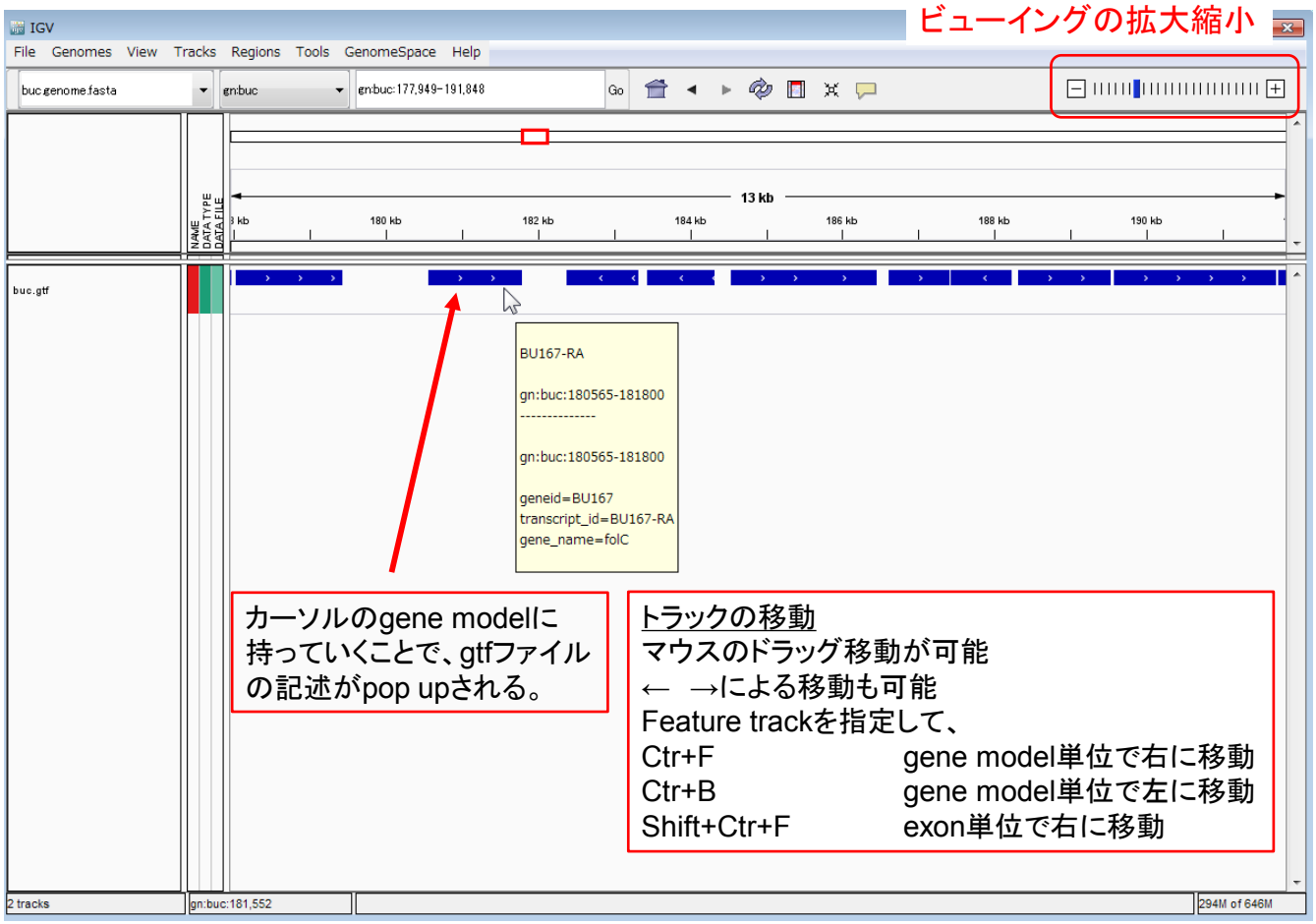
File-> Load from File
buc.gtf ファイルを読み込む

buc.gtf

gn:buc	KEGG:GENES	CDS	197	2083	.+	geneid "BU001"; transcript_id "BU001-RA"; gene_name "gidA";
gn:buc	KEGG:GENES	CDS	2278	3102	.+	geneid "BU002"; transcript_id "BU002-RA"; gene_name "atpB";
gn:buc	KEGG:GENES	CDS	3139	3378	.+	geneid "BU003"; transcript_id "BU003-RA"; gene_name "atpE";
gn:buc	KEGG:GENES	CDS	3497	3982	.+	geneid "BU004"; transcript_id "BU004-RA"; gene_name "atpF";
gn:buc	KEGG:GENES	CDS	3982	4515	.+	geneid "BU005"; transcript_id "BU005-RA"; gene_name "atpH";
gn:buc	KEGG:GENES	CDS	4530	6068	.+	geneid "BU006"; transcript_id "BU006-RA"; gene_name "atpA";
gn:buc	KEGG:GENES	CDS	6101	6973	.+	geneid "BU007"; transcript_id "BU007-RA"; gene_name "atpG";
gn:buc	KEGG:GENES	CDS	6997	8394	.+	geneid "BU008"; transcript_id "BU008-RA"; gene_name "atpD";
gn:buc	KEGG:GENES	CDS	8421	8837	.+	geneid "BU009"; transcript_id "BU009-RA"; gene_name "atpC";
gn:buc	KEGG:GENES	CDS	8911	11322	.-	geneid "BU010"; transcript_id "BU010-RA"; gene_name "gyrB";
gn:buc	KEGG:GENES	CDS	11449	12549	.-	geneid "BU011"; transcript_id "BU011-RA"; gene_name "dnaN";
gn:buc	KEGG:GENES	CDS	12554	13918	.-	geneid "BU012"; transcript_id "BU012-RA"; gene_name "dnaA";
gn:buc	KEGG:GENES	CDS	14369	14512	.+	geneid "BU013"; transcript_id "BU013-RA"; gene_name "rpmH";
gn:buc	KEGG:GENES	CDS	14525	14872	.+	geneid "BU014"; transcript_id "BU014-RA"; gene_name "rnpA";
gn:buc	KEGG:GENES	CDS	15011	16609	.+	geneid "BU015"; transcript_id "BU015-RA"; gene_name "yidC";



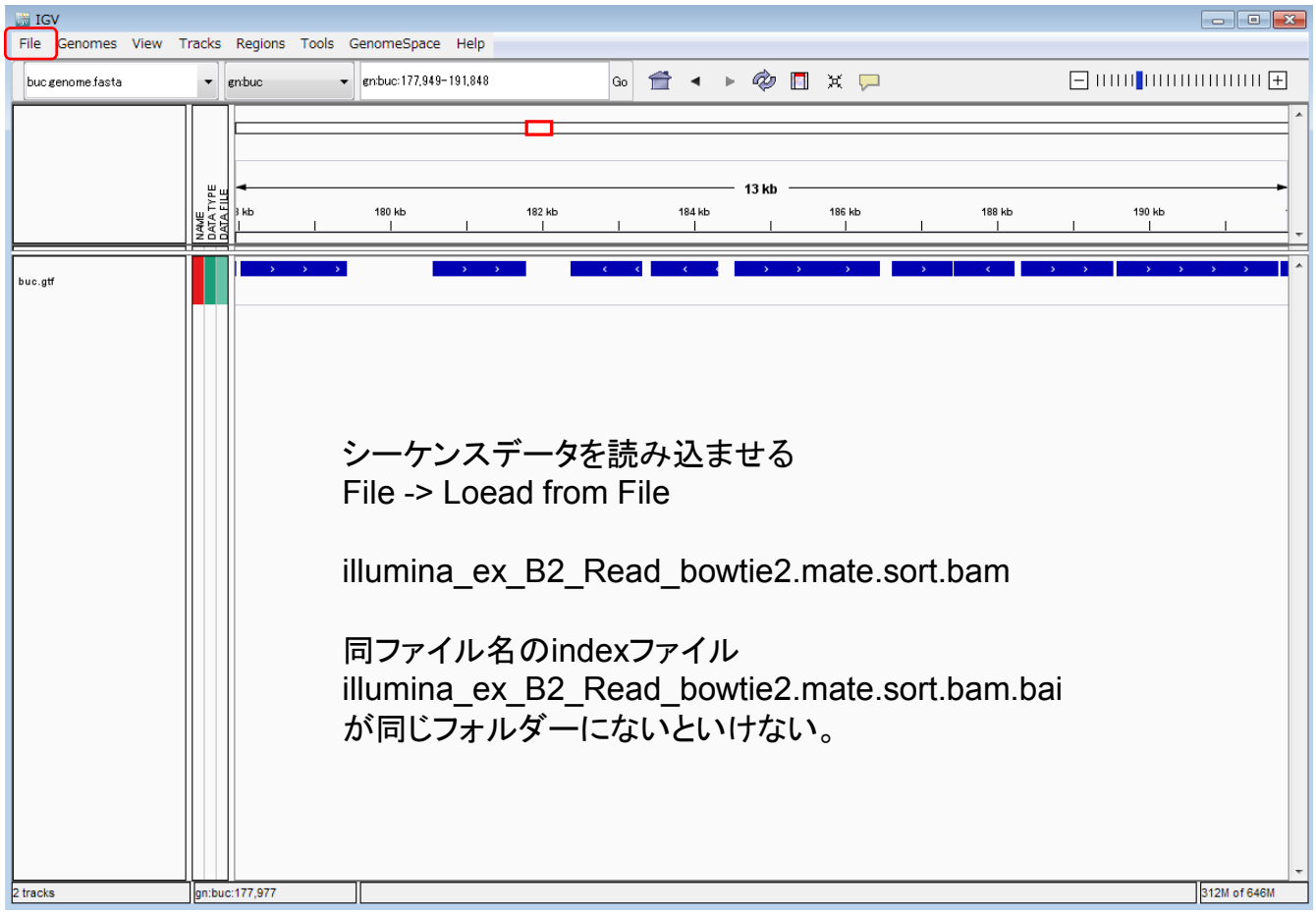
ビューイングの拡大縮小



カーソルのgene modelに持っていくことで、gtfファイルの記述がpop upされる。

トラックの移動
マウスのドラッグ移動が可能
← →による移動も可能
Feature trackを指定して、

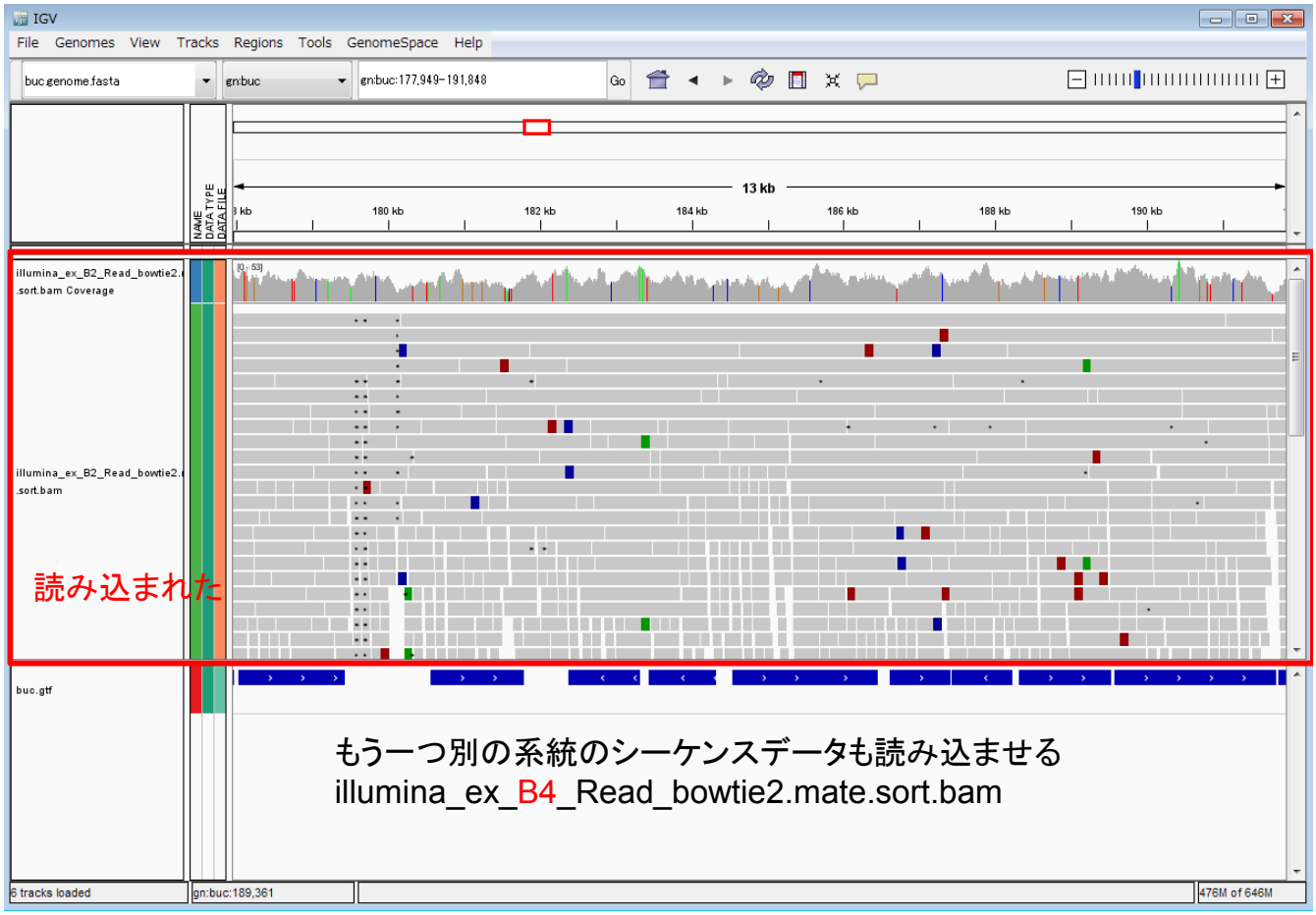
Ctr+F	gene model単位で右に移動
Ctr+B	gene model単位で左に移動
Shift+Ctr+F	exon単位で右に移動



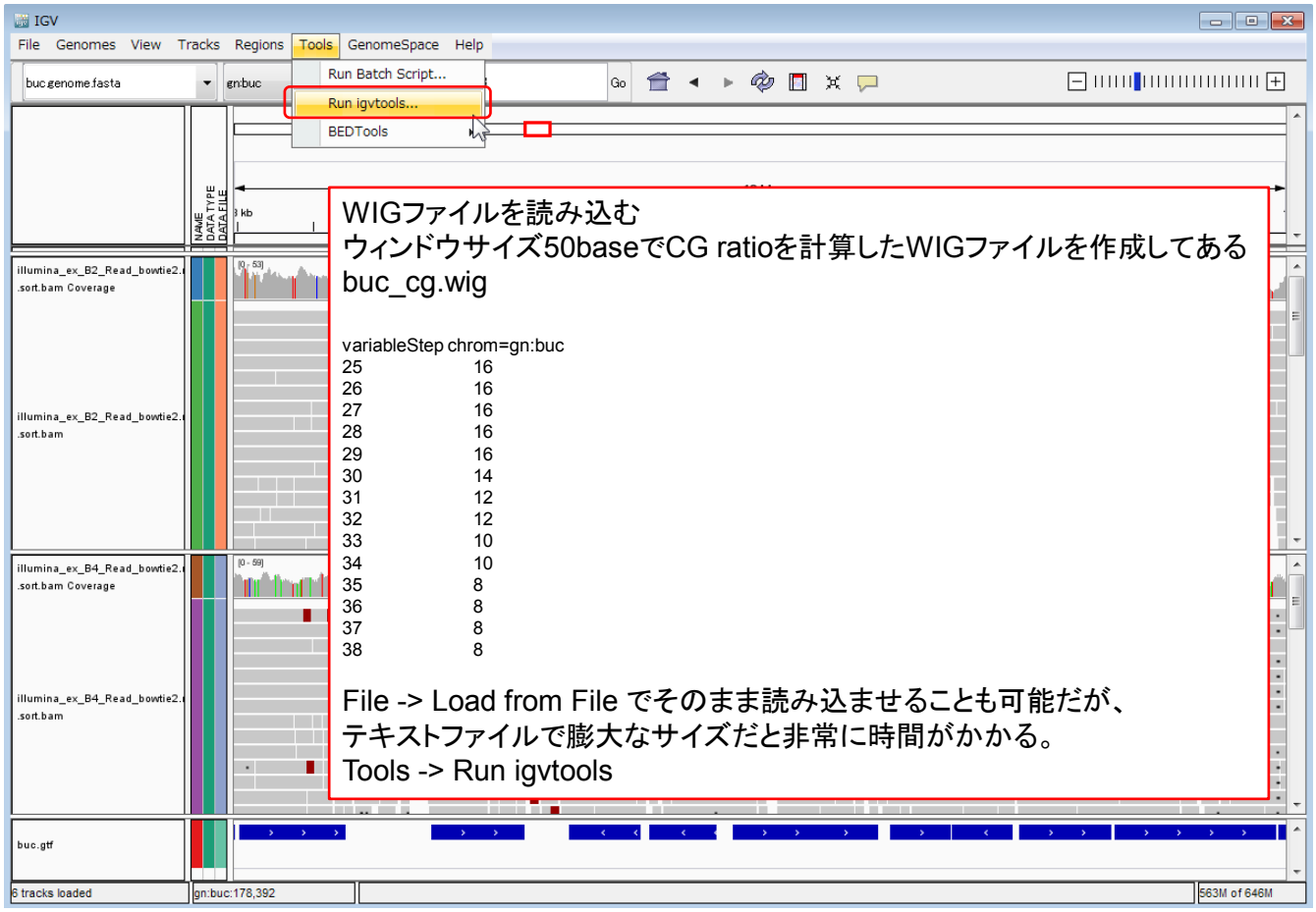
シーケンスデータを読み込ませる
File -> Loead from File

illumina_ex_B2_Read_bowtie2.mate.sort.bam

同ファイル名のindexファイル
illumina_ex_B2_Read_bowtie2.mate.sort.bam.bai
が同じフォルダーにないといけない。


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
13 kb
180 kb
182 kb
184 kb
186 kb
188 kb
190 kb
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
読み込まれた
buc.gtf
もう一つ別の系統のシーケンスデータも読み込ませる
illumina_ex_B4_Read_bowtie2.mate.sort.bam
6 tracks loaded
gn:buc:189,361
476M of 646M

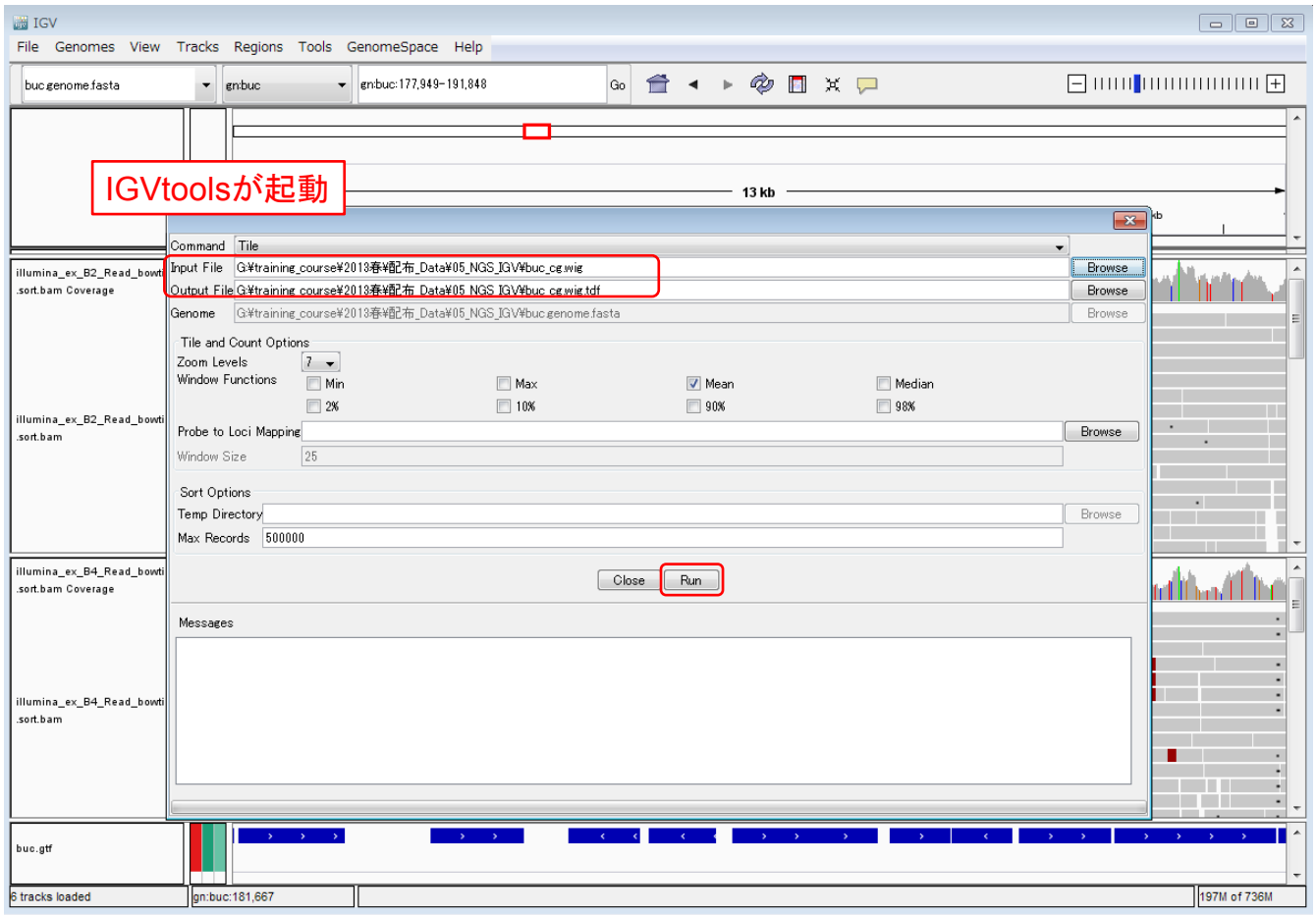

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
13 kb
180 kb
182 kb
184 kb
186 kb
188 kb
190 kb
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
6 tracks loaded
gn:buc:179,350
545M of 646M
こういう所を
クリックしたまま
移動させることで
幅を変更
各Trackはドラッグアンドドロップで移動できる

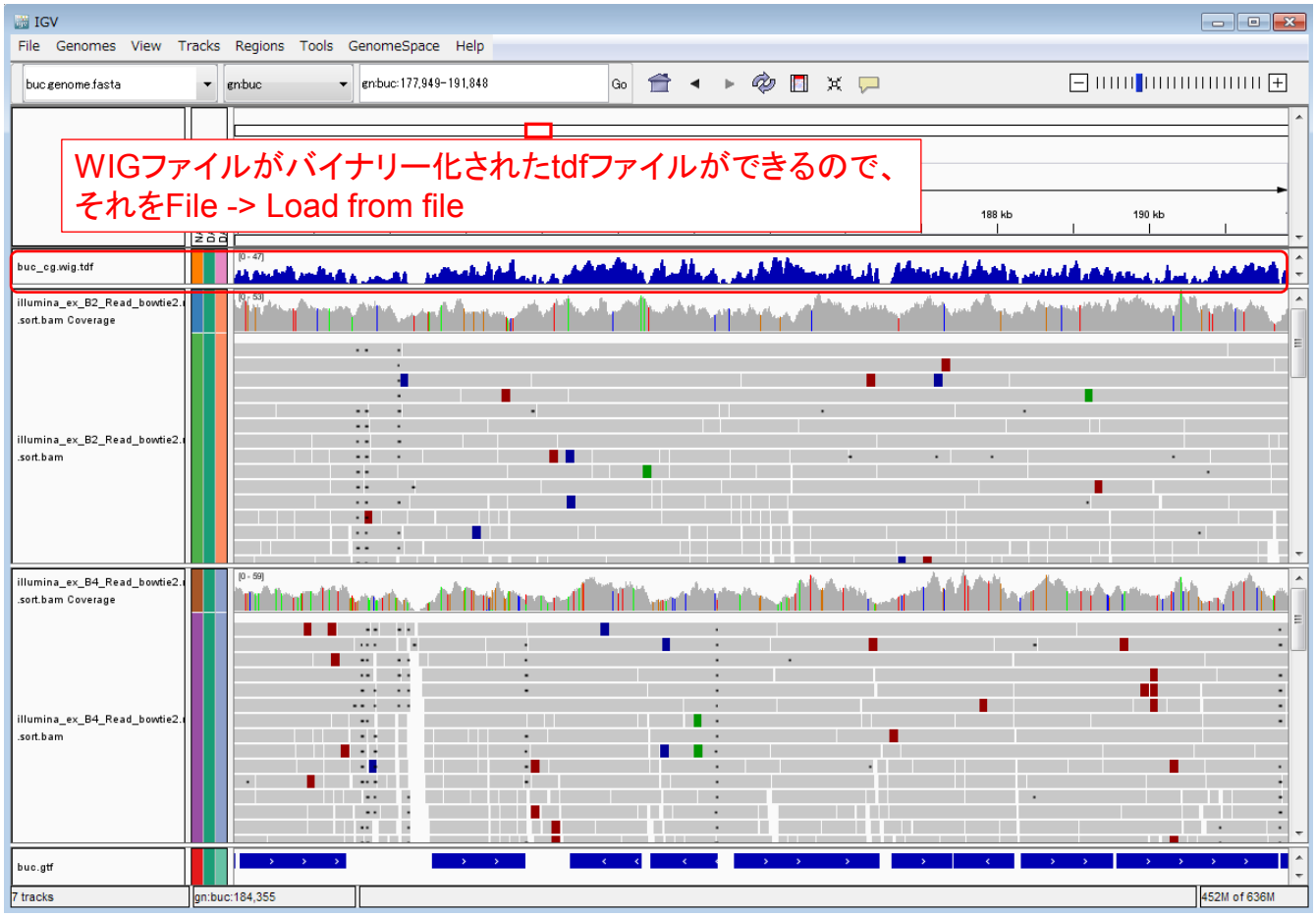


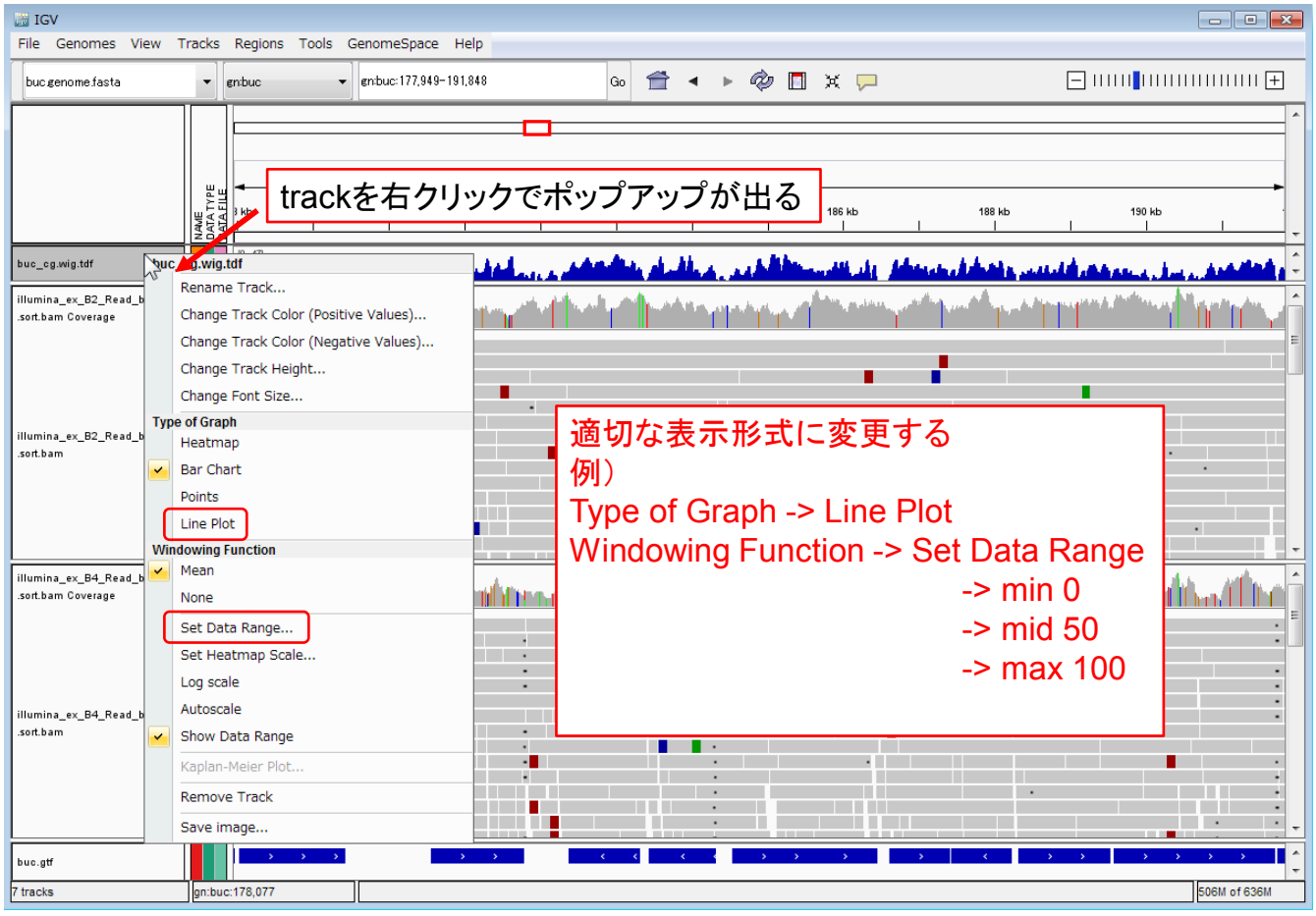
WIGファイルを読み込む
ウィンドウサイズ50baseでCG ratioを計算したWIGファイルを作成してある
buc_cg.wig

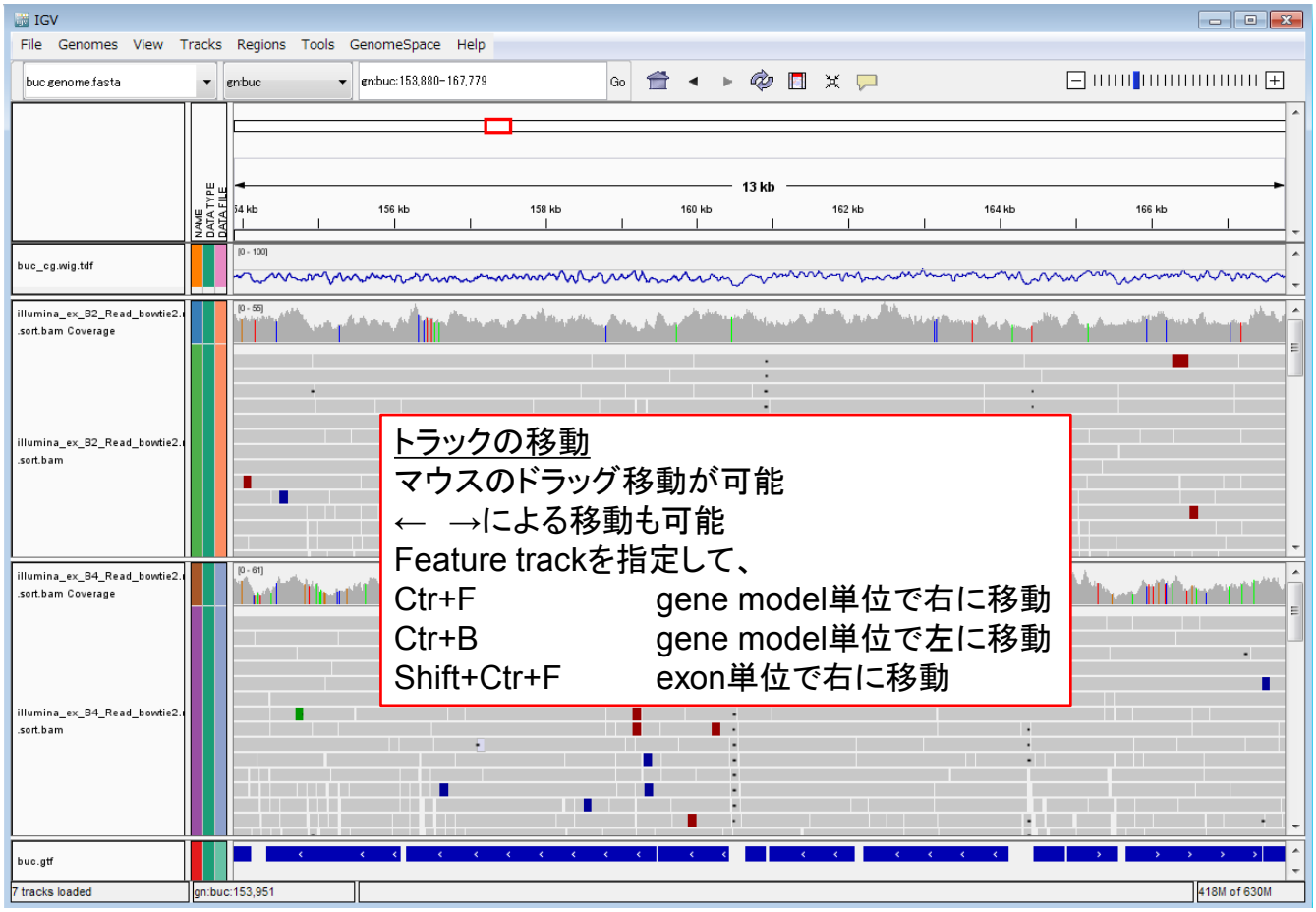
```
variableStep chrom=gn:buc
25              16
26              16
27              16
28              16
29              16
30              14
31              12
32              12
33              10
34              10
35              8
36              8
37              8
38              8
```

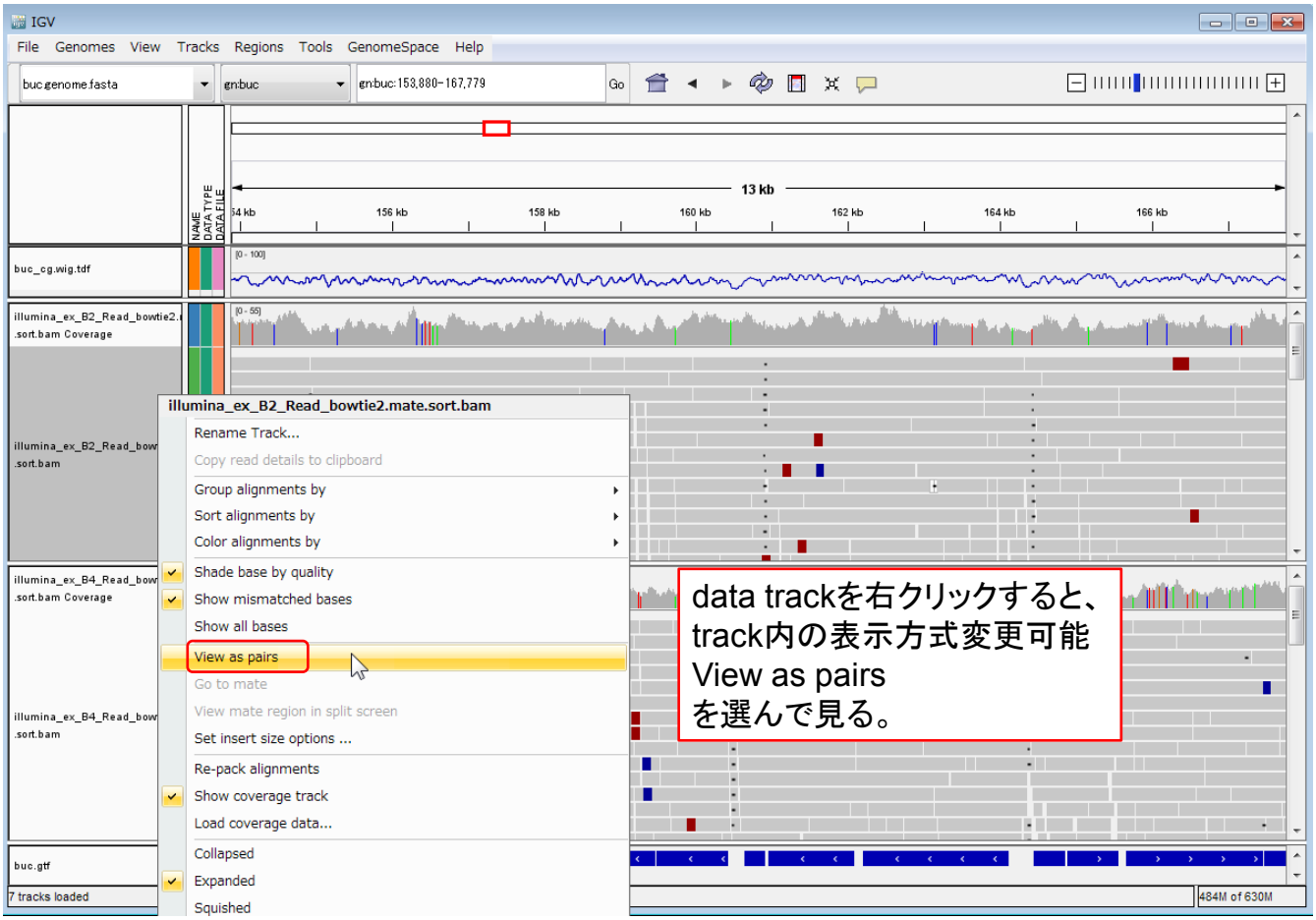
File -> Load from File でそのまま読み込ませることも可能だが、
テキストファイルで膨大なサイズだと非常に時間がかかる。
Tools -> Run igvtools

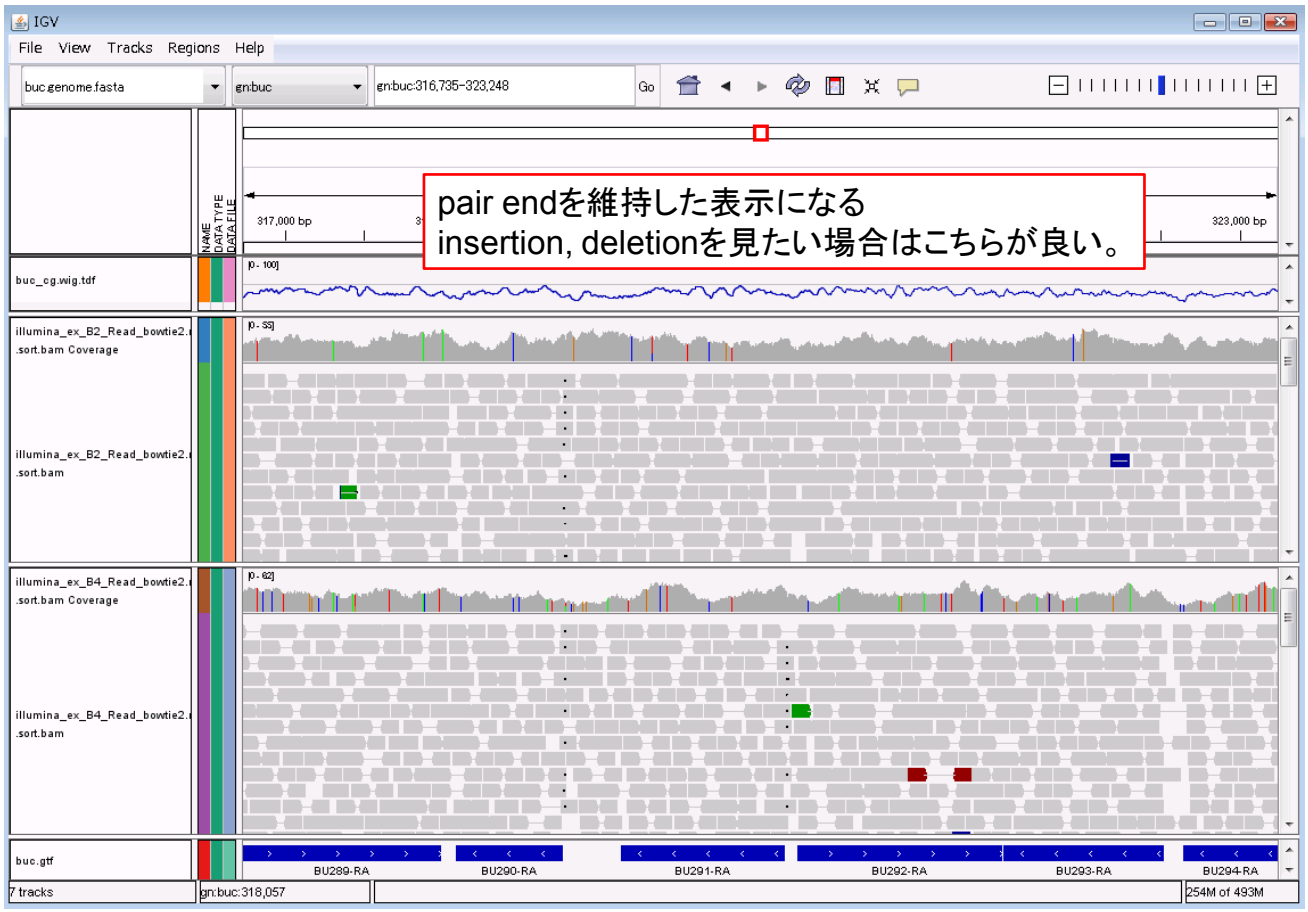


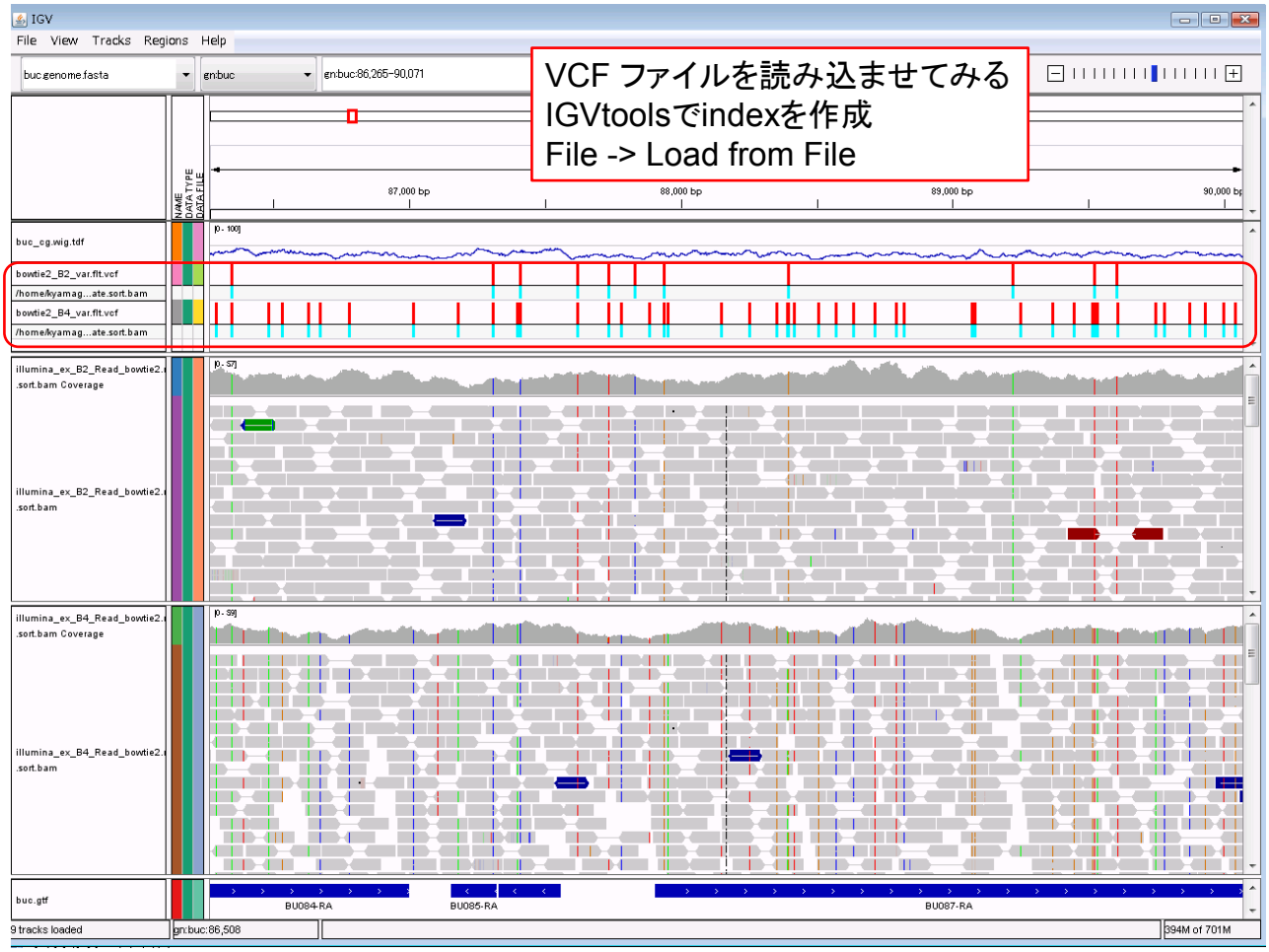

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
WIGファイルがバイナリー化されたtdfファイルができるので、
それをFile -> Load from file
188 kb
190 kb
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
7 tracks
gn:buc:184,355
452M of 636M

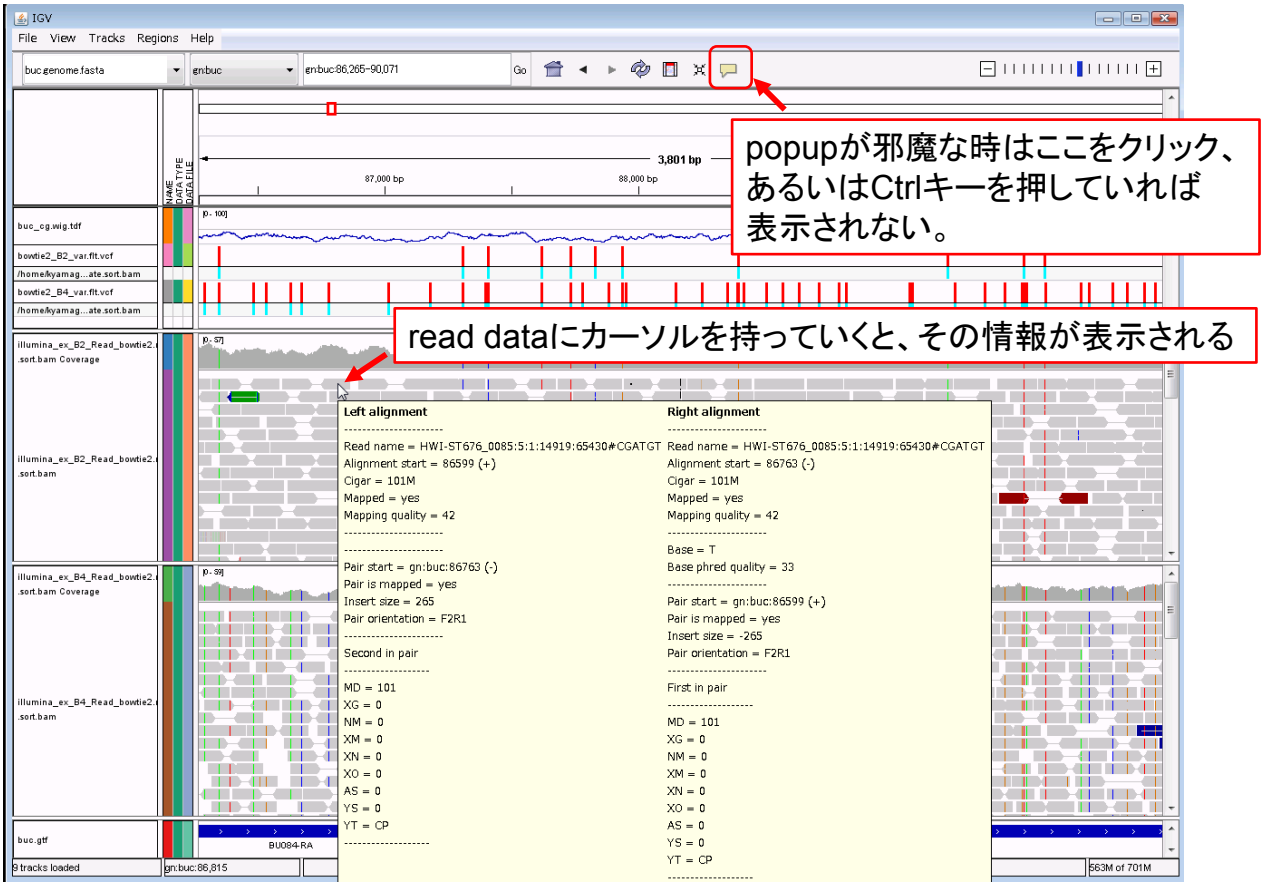

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
trackを右クリックでポップアップが出る
186 kb
188 kb
190 kb
buc_cg.wig.tdf
Rename Track...
Change Track Color (Positive Values)...
Change Track Color (Negative Values)...
Change Track Height...
Change Font Size...
Type of Graph
Heatmap
Bar Chart
Points
Line Plot
Windowing Function
Mean
None
Set Data Range...
Set Heatmap Scale...
Log scale
Autoscale
Show Data Range
Kaplan-Meier Plot...
Remove Track
Save image...
適切な表示形式に変更する
例）
Type of Graph -> Line Plot
Windowing Function -> Set Data Range
-> min 0
-> mid 50
-> max 100
buc.gtf
7 tracks
gn:buc:178,077
506M of 636M


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:153,880-167,779
Go
13 kb
154 kb
156 kb
158 kb
160 kb
162 kb
164 kb
166 kb
NAME
DATA TYPE
DATA FILE
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
7 tracks loaded
gn:buc:153,951
418M of 630M
トラックの移動
マウスのドラッグ移動が可能
← →による移動も可能
Feature trackを指定して、
Ctr+F gene model単位で右に移動
Ctr+B gene model単位で左に移動
Shift+Ctr+F exon単位で右に移動


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:153,880-167,779
Go
13 kb
154 kb
156 kb
158 kb
160 kb
162 kb
164 kb
166 kb
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.mate.sort.bam
Rename Track...
Copy read details to clipboard
Group alignments by
Sort alignments by
Color alignments by
Shade base by quality
Show mismatched bases
Show all bases
View as pairs
Go to mate
View mate region in split screen
Set insert size options ...
Re-pack alignments
Show coverage track
Load coverage data...
Collapsed
Expanded
Squished
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bow
.sort.bam
illumina_ex_B4_Read_bow
.sort.bam Coverage
illumina_ex_B4_Read_bow
.sort.bam
buc.gtf
7 tracks loaded
484M of 630M
data trackを右クリックすると、
track内の表示方式変更可能
View as pairs
を選んで見る。


IGV
File View Tracks Regions Help
buc.genome.fasta
gn:buc
gn:buc:316,735-323,248
Go
317,000 bp
323,000 bp
pair endを維持した表示になる
insertion, deletionを見たい場合はこちらが良い。
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
BU289-RA
BU290-RA
BU291-RA
BU292-RA
BU293-RA
BU294-RA
7 tracks
gn:buc:318,057
254M of 493M


IGV
File View Tracks Regions Help
buc.genome.fasta
gn:buc
gn:buc:86,265-90,071
VCF ファイルを読み込ませてみる
IGVtoolsでindexを作成
File -> Load from File
87,000 bp
88,000 bp
89,000 bp
90,000 bp
buc_cg.wig.tdf
bowtie2_B2_var.flt.vcf
/home/kyamag...ate.sort.bam
bowtie2_B4_var.flt.vcf
/home/kyamag...ate.sort.bam
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
BU084-RA
BU085-RA
BU087-RA
9 tracks loaded
gn:buc:86,508
394M of 701M



IGV紹介のまとめ

可視化ツールとして十分な機能を持つ

- 無料
- 比較的簡単・お手軽
- 自分で見るためにも良し、人に見せるためにも良し
- 利用範囲は次世代DNAシーケンサーに限定しない
  広くゲノミクスの解析に有用

ごく一部のみの機能を紹介しました。
ウェブサイトを見ながら復習をお勧めします。