# 第5回バイオインフォマティクス実習コース

横浜市大　先端医科学研究センター
バイオインフォマティクス研究室
室長　田村智彦
准教授　中林潤
免疫学　藩龍馬

- RNA-seqデータ解析

# RNA-seqデータ解析の手順

- シーケンス
- ゲノム上にマッピング
- 発現量に換算
- 発現解析

# マッピングツール
# tophat

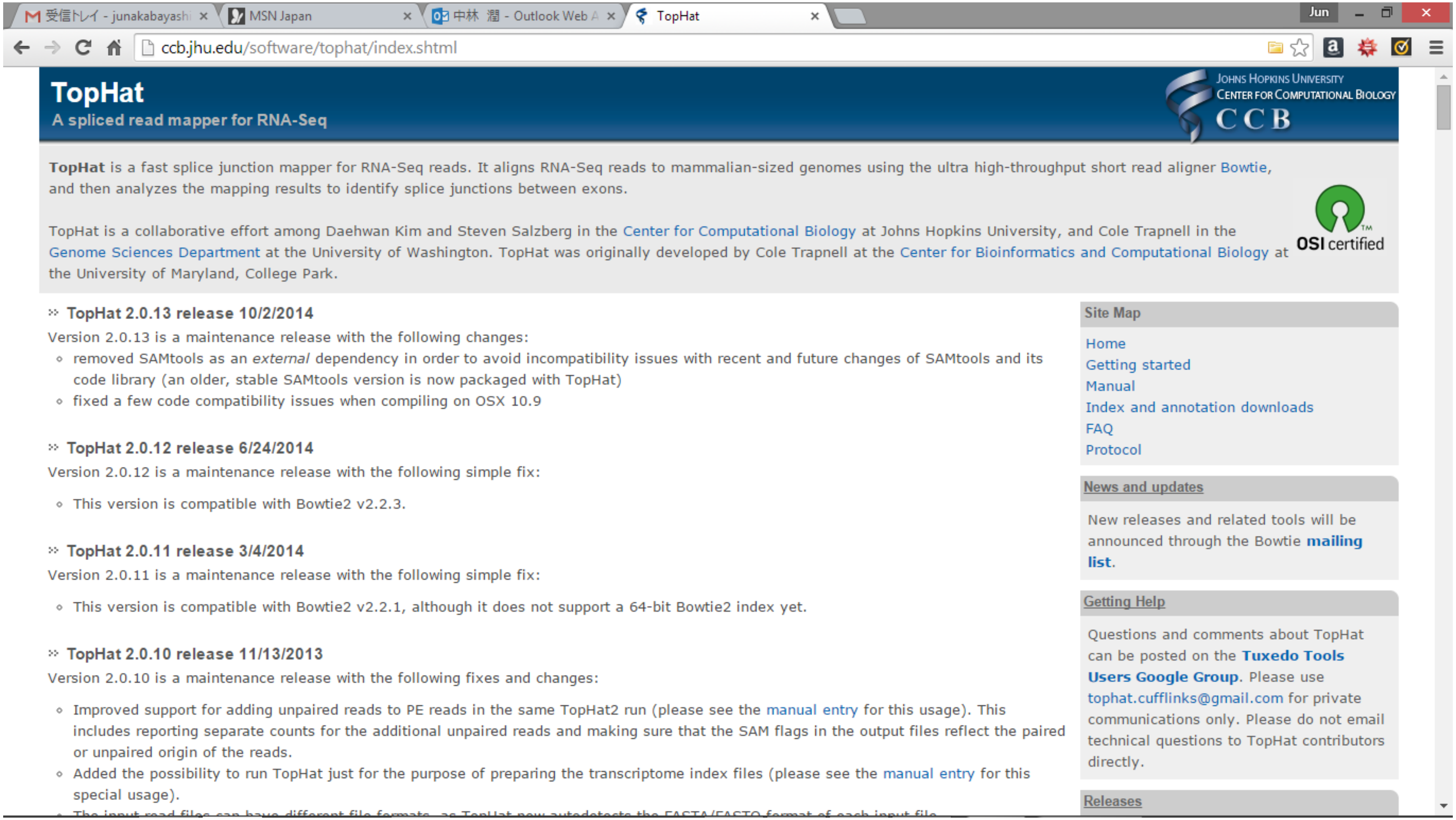
- Johns Hopkins University
  Center for Computational Biology

- http://ccb.jhu.edu/software/tophat/index.shtml

- Transcriptome解析用マッピングツール
  Bowtie2を呼び出してマッピング
  スプライスジャンクションを予測する

# TopHat

http://ccb.jhu.edu/software/tophat/index.shtml


受信トレイ - junakabayashi
MSN Japan
中林 潤 - Outlook Web A
TopHat
Jun
ccb.jhu.edu/software/tophat/index.shtml
Johns Hopkins University
Center for Computational Biology
CCB
OSI certified

**TopHat**

A spliced read mapper for RNA-Seq

**TopHat** is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

**TopHat 2.0.13 release 10/2/2014**

Version 2.0.13 is a maintenance release with the following changes:

- removed SAMtools as an *external* dependency in order to avoid incompatibility issues with recent and future changes of SAMtools and its code library (an older, stable SAMtools version is now packaged with TopHat)
- fixed a few code compatibility issues when compiling on OSX 10.9

**TopHat 2.0.12 release 6/24/2014**

Version 2.0.12 is a maintenance release with the following simple fix:

- This version is compatible with Bowtie2 v2.2.3.

**TopHat 2.0.11 release 3/4/2014**

Version 2.0.11 is a maintenance release with the following simple fix:

- This version is compatible with Bowtie2 v2.2.1, although it does not support a 64-bit Bowtie2 index yet.

**TopHat 2.0.10 release 11/13/2013**

Version 2.0.10 is a maintenance release with the following fixes and changes:

- Improved support for adding unpaired reads to PE reads in the same TopHat2 run (please see the manual entry for this usage). This includes reporting separate counts for the additional unpaired reads and making sure that the SAM flags in the output files reflect the paired or unpaired origin of the reads.
- Added the possibility to run TopHat just for the purpose of preparing the transcriptome index files (please see the manual entry for this special usage).
- The input read files can have different file formats, as TopHat now autodetects the FASTA/FASTQ format of each input file

**Site Map**

- Home
- Getting started
- Manual
- Index and annotation downloads
- FAQ
- Protocol

**News and updates**

New releases and related tools will be announced through the Bowtie **mailing list**.

**Getting Help**

Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

**Releases**

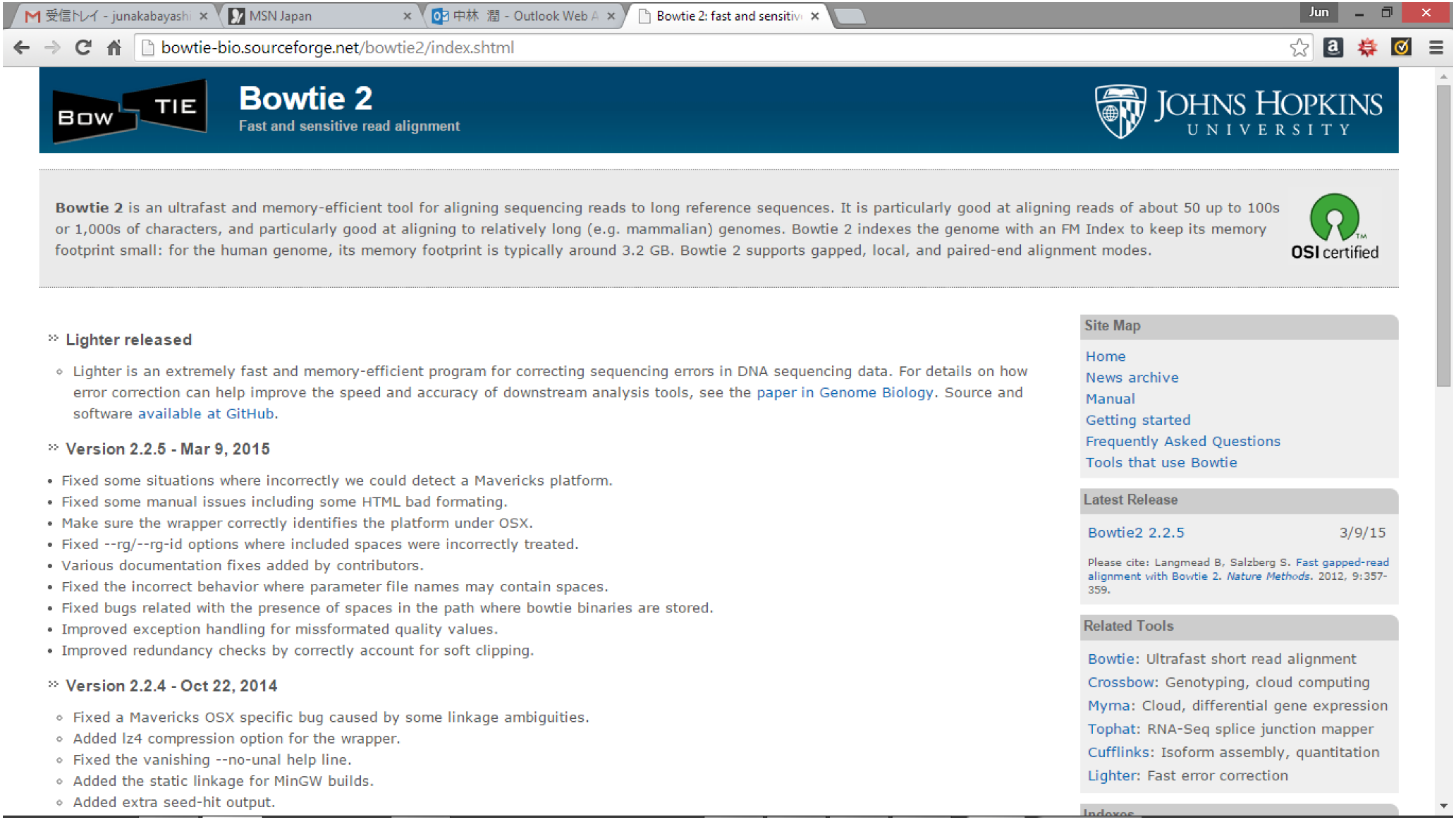
# マッピングツール
# Bowtie2

- John Hopkins University

- http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

# Bowtie2

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml



受信トレイ - junakabayashi
MSN Japan
中林 潤 - Outlook Web A
Bowtie 2: fast and sensitiv
Jun
bowtie-bio.sourceforge.net/bowtie2/index.shtml
BOW TIE
JOHNS HOPKINS UNIVERSITY
OSI certified

## Bowtie 2

Fast and sensitive read alignment

**Bowtie 2** is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

### Lighter released

- Lighter is an extremely fast and memory-efficient program for correcting sequencing errors in DNA sequencing data. For details on how error correction can help improve the speed and accuracy of downstream analysis tools, see the paper in Genome Biology. Source and software available at GitHub.

### Version 2.2.5 - Mar 9, 2015

- Fixed some situations where incorrectly we could detect a Mavericks platform.
- Fixed some manual issues including some HTML bad formating.
- Make sure the wrapper correctly identifies the platform under OSX.
- Fixed --rg/--rg-id options where included spaces were incorrectly treated.
- Various documentation fixes added by contributors.
- Fixed the incorrect behavior where parameter file names may contain spaces.
- Fixed bugs related with the presence of spaces in the path where bowtie binaries are stored.
- Improved exception handling for missformated quality values.
- Improved redundancy checks by correctly account for soft clipping.

### Version 2.2.4 - Oct 22, 2014

- Fixed a Mavericks OSX specific bug caused by some linkage ambiguities.
- Added lz4 compression option for the wrapper.
- Fixed the vanishing --no-unal help line.
- Added the static linkage for MinGW builds.
- Added extra seed-hit output.

**Site Map**

Home
News archive
Manual
Getting started
Frequently Asked Questions
Tools that use Bowtie

**Latest Release**

Bowtie2 2.2.5 3/9/15

Please cite: Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 2012, 9:357-359.

**Related Tools**

Bowtie: Ultrafast short read alignment
Crossbow: Genotyping, cloud computing
Myrna: Cloud, differential gene expression
Tophat: RNA-Seq splice junction mapper
Cufflinks: Isoform assembly, quantitation
Lighter: Fast error correction

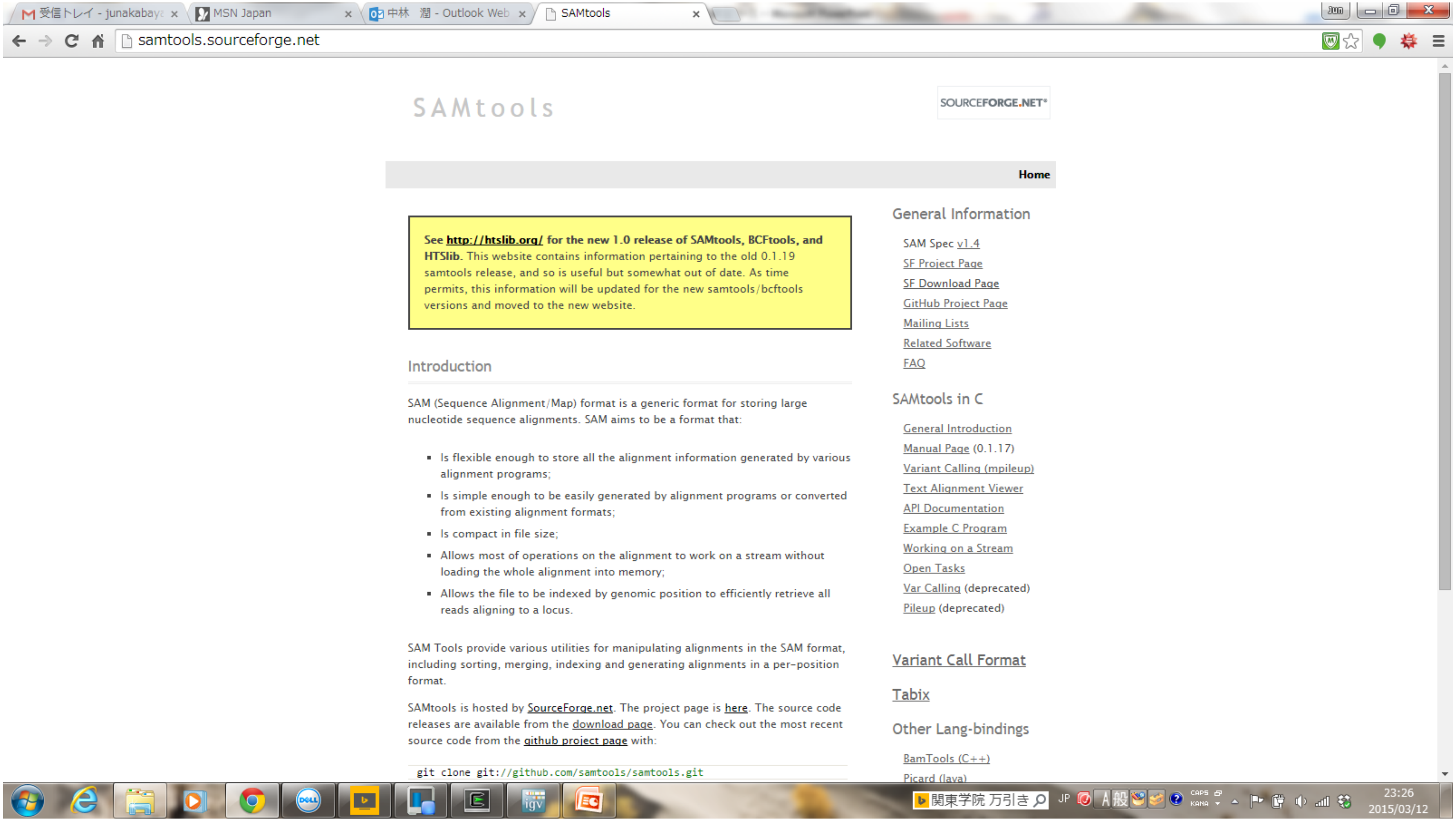
**Indexes**

# samtools

- http://samtools.sourceforge.net/
- sam→bam変換
- sam fileのsort
- index作成

# SAMtools

http://samtools.sourceforge.net/



SAMtools

**See http://htslib.org/ for the new 1.0 release of SAMtools, BCFtools, and HTSlib.** This website contains information pertaining to the old 0.1.19 samtools release, and so is useful but somewhat out of date. As time permits, this information will be updated for the new samtools/bcftools versions and moved to the new website.

## Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

SAMtools is hosted by SourceForge.net. The project page is here. The source code releases are available from the download page. You can check out the most recent source code from the github project page with:

```
git clone git://github.com/samtools/samtools.git
```

## General Information

- SAM Spec v1.4
- SF Project Page
- SF Download Page
- GitHub Project Page
- Mailing Lists
- Related Software
- FAQ

## SAMtools in C

- General Introduction
- Manual Page (0.1.17)
- Variant Calling (mpileup)
- Text Alignment Viewer
- API Documentation
- Example C Program
- Working on a Stream
- Open Tasks
- Var Calling (deprecated)
- Pileup (deprecated)

## Variant Call Format

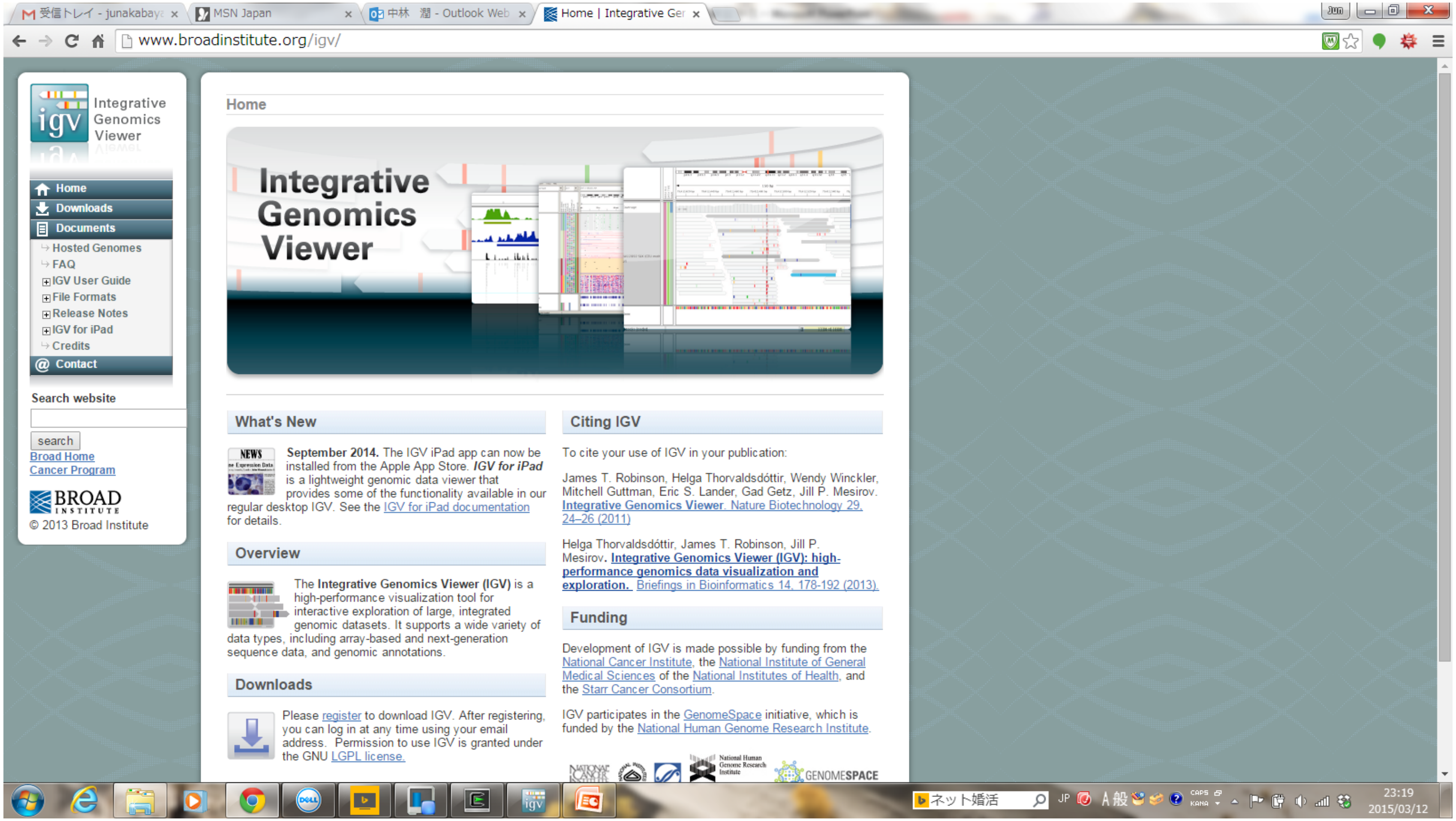
## Tabix

## Other Lang-bindings

- BamTools (C++)
- Picard (Java)

# Integrative Genomics Viewer

Broad institute
http://broadinstitute.org/igv/


受信トレイ - junakabaya
MSN Japan
中林　潤 - Outlook Web
Home | Integrative Ger
www.broadinstitute.org/igv/
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
NEWS
National Human Genome Research Institute
GENOMESPACE
ネット婚活
JP
23:19
2015/03/12

## 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.

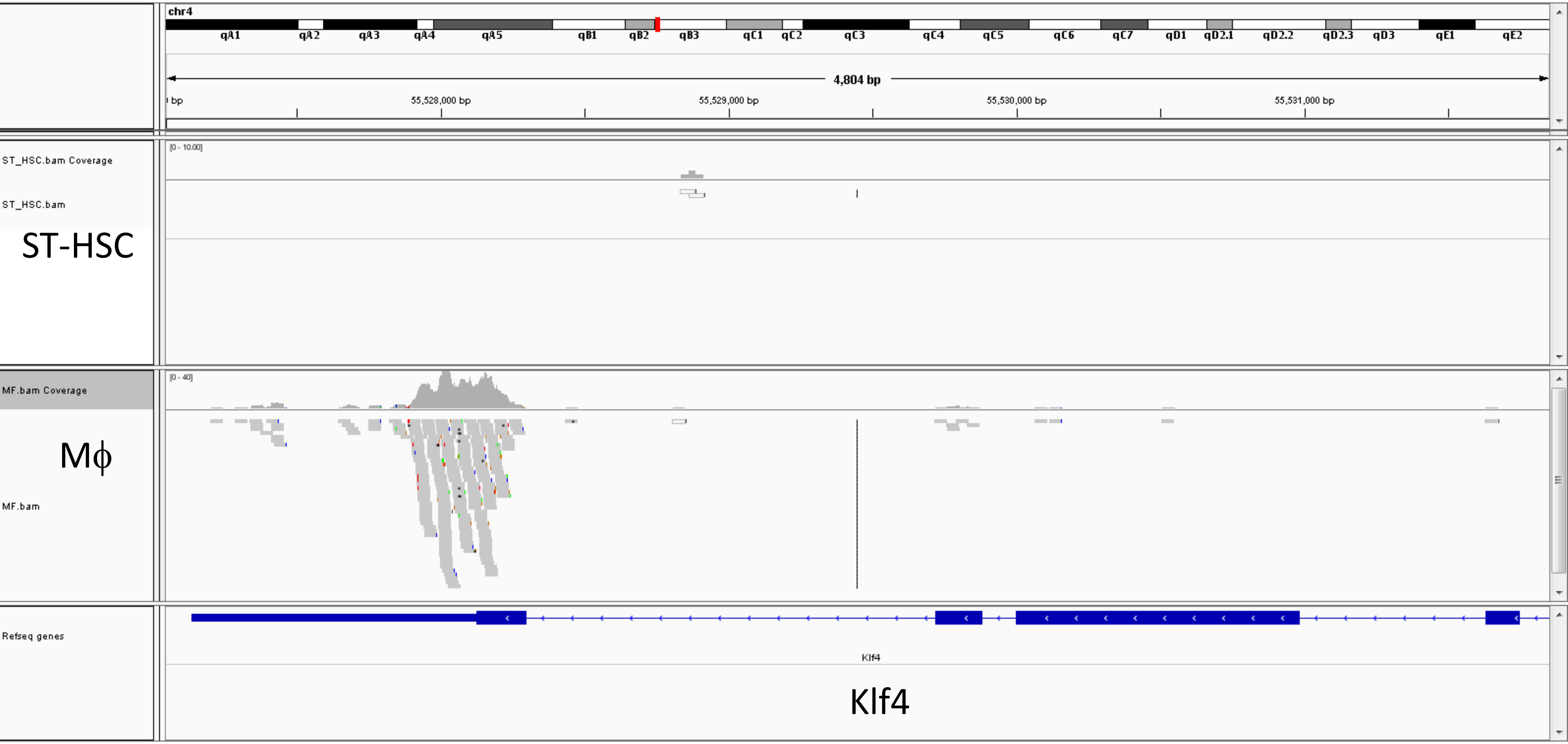
IGV participates in the GenomeSpace initiative, which is funded by the National Human Genome Research Institute.

GSE60101から
ST_HSC、Mϕの
遺伝子発現プロファイル
FASTQ fileを取得

tophatで
マッピング

bam file
samtoolsで
index作成

integrative genomics viewerで
表示


chr4
qA1
qA2
qA3
qA4
qA5
qB1
qB2
qB3
qC1
qC2
qC3
qC4
qC5
qC6
qC7
qD1
qD2.1
qD2.2
qD2.3
qD3
qE1
qE2
4,804 bp
55,528,000 bp
55,529,000 bp
55,530,000 bp
55,531,000 bp
ST_HSC.bam Coverage
ST_HSC.bam
[0 - 10.00]
ST-HSC
MF.bam Coverage
MF.bam
[0 - 40]
Mϕ
Refseq genes
Klf4
Klf4

# cufflinks
# 発現定量

- マッピングデータを発現量に換算

⬇

RPKM

reads per kilobase of exon

per million mapped sequence reads

マッピングされたリード数をエクソン長と総リード数で正規化した値

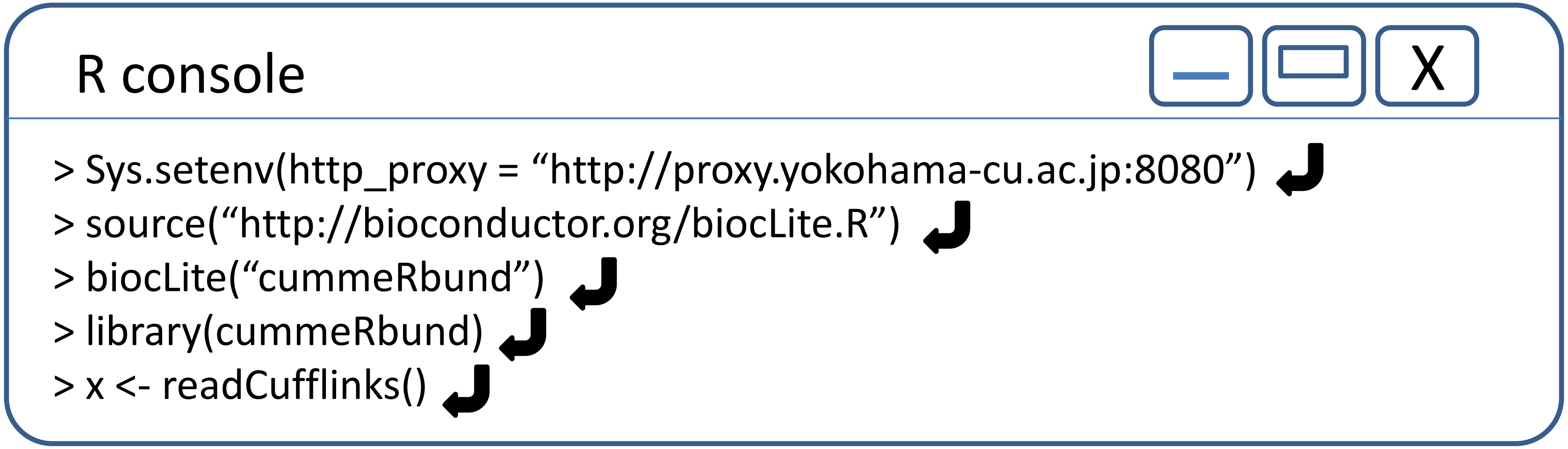
$$RPKM = \frac{X_t}{l_t N} \times 10^9$$

**$X_t$** : 転写物tにマップされたリード数

**$l_t$** : 転写物tの長さ

**$N$** : 総リード数

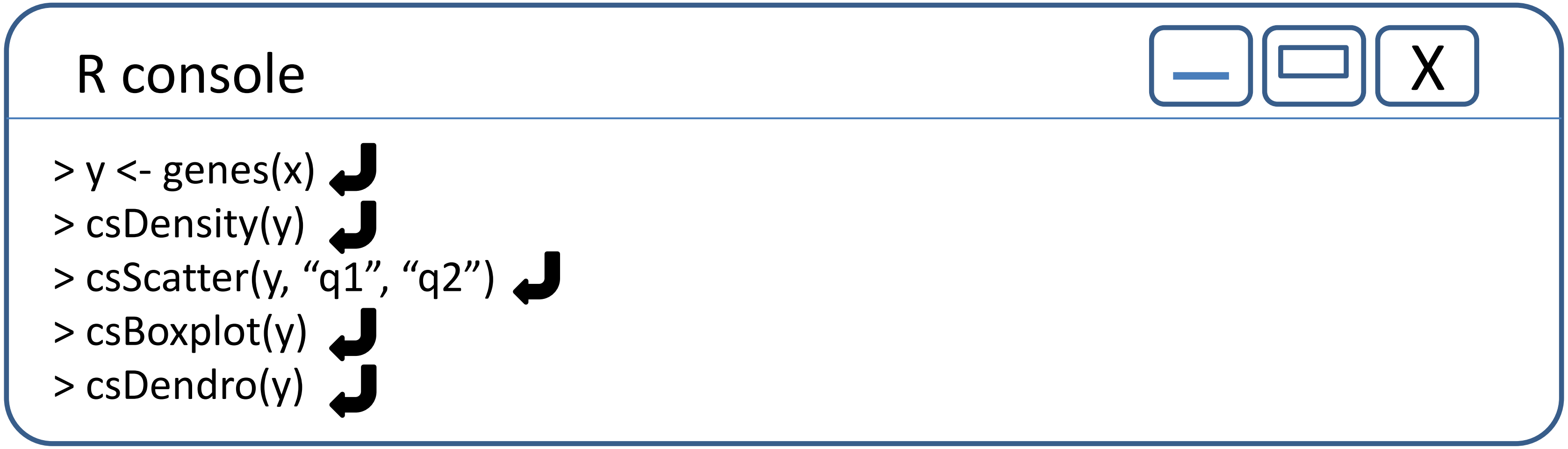
# R package “cummeRbund”



```
> Sys.setenv(http_proxy = “http://proxy.yokohama-cu.ac.jp:8080”) ↵
> source(“http://bioconductor.org/biocLite.R”) ↵
> biocLite(“cummeRbund”) ↵
> library(cummeRbund) ↵
> x <- readCufflinks() ↵
```

- proxyの設定
- biocLite.Rの設定
- パッケージ“cummeRbund”の読み込み
- 変数xに発現量データ（cuffdiffの出力）を格納

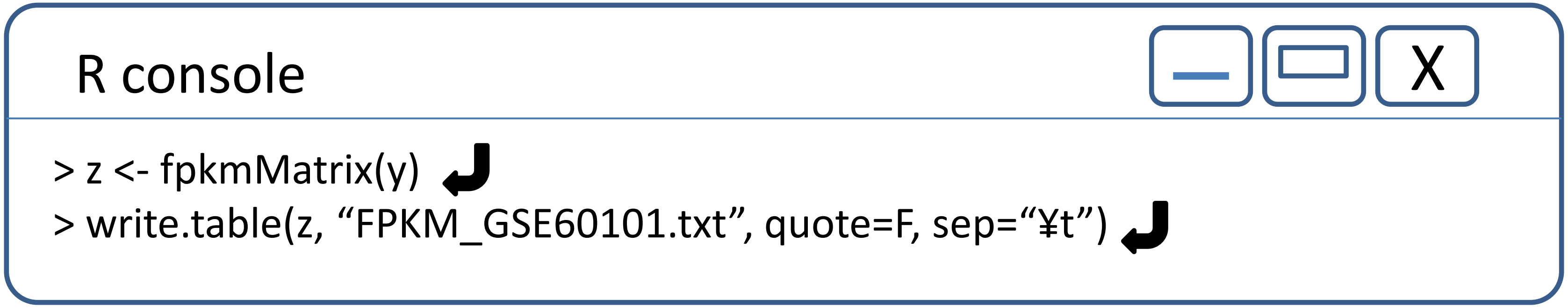
# R package “cummeRbund”



```
> y <- genes(x) ↵
> csDensity(y) ↵
> csScatter(y, “q1”, “q2”) ↵
> csBoxplot(y) ↵
> csDendro(y) ↵
```

- 遺伝子ごとの発現量を取得し、変数yに格納
- density plot, dendrogram, scatter plot, boxplot, dendrgramを作図

# R package “cummeRbund”



```
> z <- fpkmMatrix(y) ↵
> write.table(z, “FPKM_GSE60101.txt”, quote=F, sep=“¥t”) ↵
```

- 発現量データを変数zに格納
- タブ区切りテキストファイルとして出力

アンケートにご協力をお願いいたします。