

# 次世代シーケンスデータベース

Prepared for AJACS79 (AJACS十勝2)

12 Sep 2019

大田達郎

Tazro Ohta

[t.ohta@dbcls.rois.ac.jp](mailto:t.ohta@dbcls.rois.ac.jp)

情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター

Database Center for Life Science (DBCLS), Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS)

これはバイオサイエンスデータベースセンター (NBDC) が主催する統合データベース講習会 AJACS79 (AJACS十勝2) 「次世代シーケンスデータベース」の資料です。オンラインで閲覧することを想定しており、テキスト中にハイパーリンクが埋め込まれている箇所があります。また、随時アップデートをしていますので、最新版は <https://github.com/AJACS-training/AJACS79> からご覧ください。

講習会のお知らせとプログラム、各講習の資料へのリンクはこちら <https://biosciencedbc.jp/event/ajacs/ajacs79.html> です。

## 概要

本講習では、主に新型DNAシーケンサー (High-Throughput Sequencer, HTS) から得られる塩基配列データと、それに基づく生物学的データを公開しているデータベースの概要、およびデータベースからのデータの取得の手順を学びます。

## 講習の流れ

今回の講習では、以下の内容について説明します。

- 新型シーケンサーとは
  - 定義
  - どのように使われているか
  - 機器について
  - シーケンシング・アプリケーションについて
  - 得られるデータについて
- 新型シーケンサーのデータベースとは
  - データが登録されるまで
  - 一次データレポジトリ
  - 二次データを含むデータベース
- データ解析についてのTips

## 新型シーケンサーとは

### 定義

講習のタイトルには「次世代シーケンサー」と書いていますが、現存するので「次世代」ではありません。「超並列シーケンス」、「新型シーケンス」とも呼ばれます。「次世代」は、いわゆる [バズワード](#) ので、明確な定義は存在しませんが、「サンガー法以降、2000年代中頃から登場した、新しいDNAシーケンス技術の総称」という理解で問題ありません。新型シーケンサーとして含まれるのは

- Illumina
- IonTorrent
- Roche 454
- PacBio RS
- SOLiD
- Oxford Nanopore

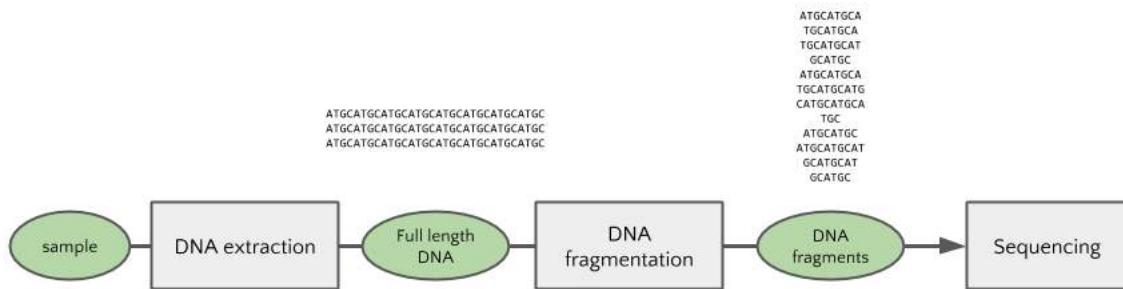
などのシーケンサーベンダー・ブランドによって提供する技術に代表されます。これまでに登場した新型シーケンサーの測定技術の詳細については、次の文献を参照してください。

[Goodwin, Sara, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies." \*Nature Reviews Genetics\* 17.6 \(2016\): 333-351.](#)

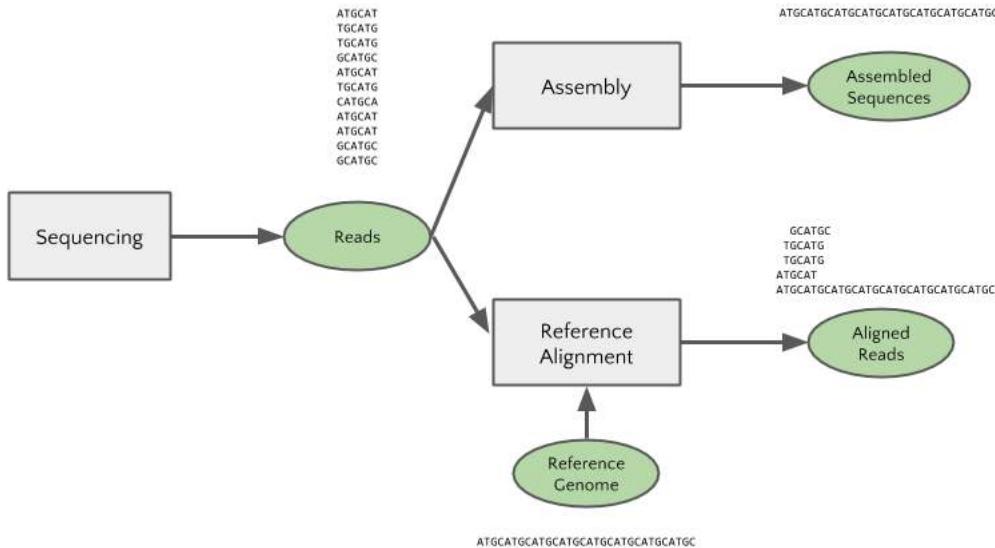
## どのように使われているか



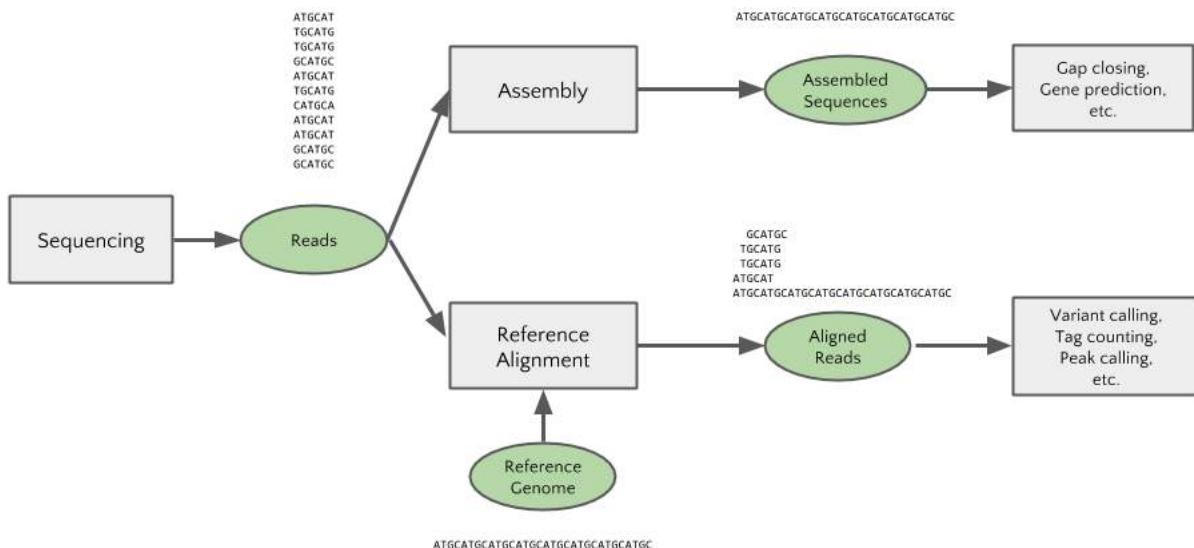
新型シーケンサはサンプルDNAを入力として処理し、DNA塩基配列をデータとして出力します。新型シーケンサはサンガー法のように長い塩基配列を読めない代わりに、大量の塩基配列を並列に読むことで高いスループットを実現します。ほとんどの場合、抽出されたDNAは断片化され、シーケンス反応用のプライマーを処理したのちにシーケンサーにかけられます。



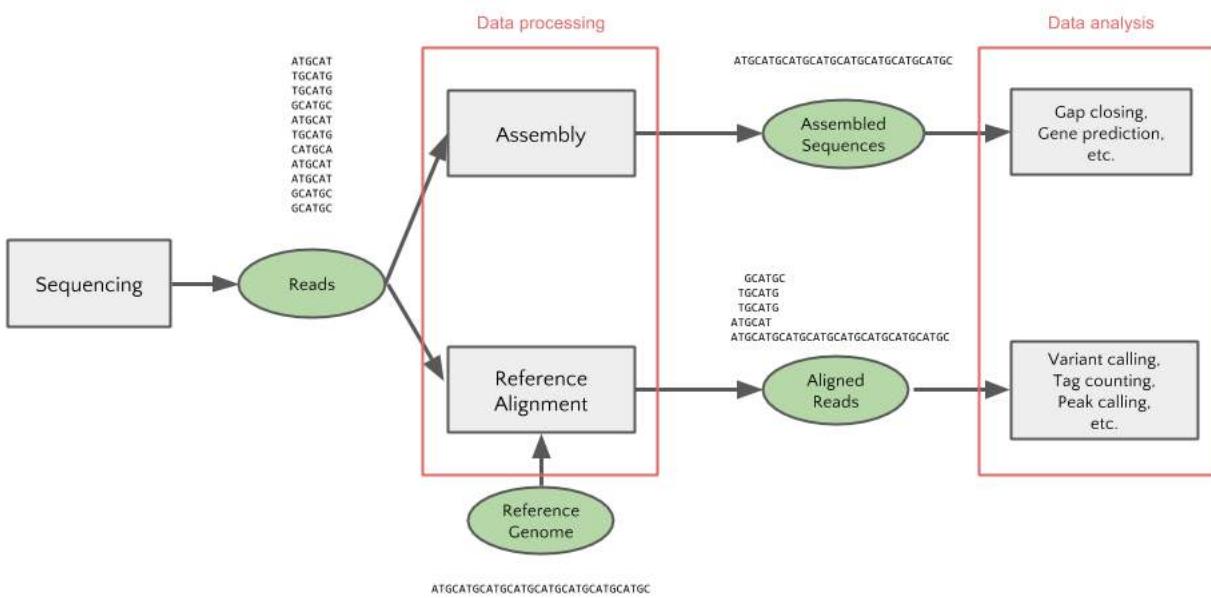
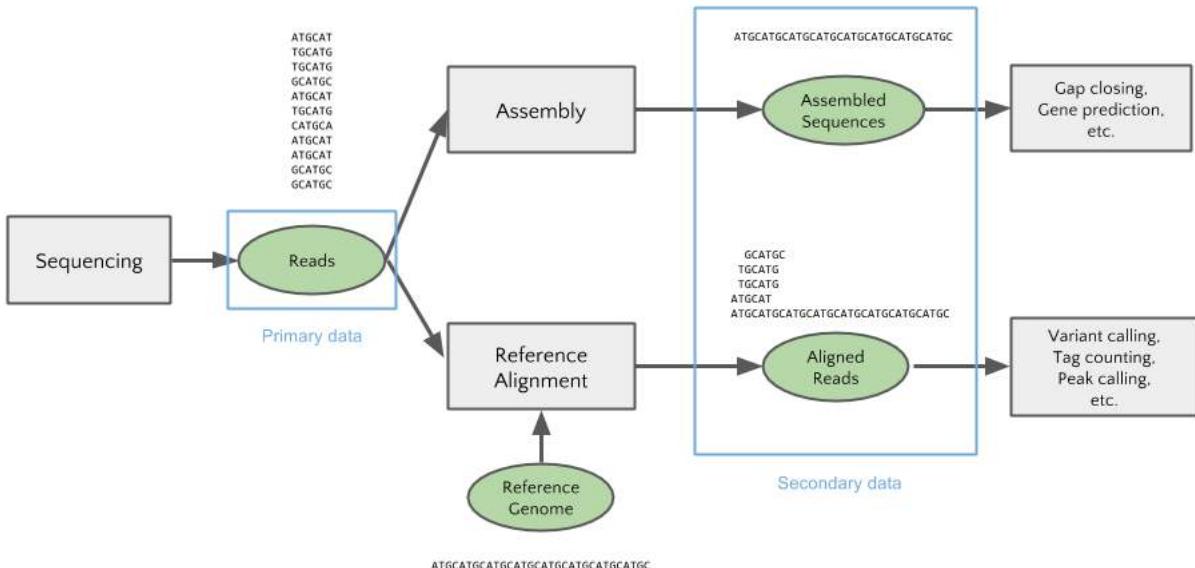
シーケンサーから出力される一次データ (primary data, 生データ) はこの断片化された配列 (リード) の情報です。そのため、出力されたデータは、そのままでは生物学的な解釈ができません。そこで、新型シーケンサのデータを研究に利用するためには、リードを元の塩基配列に復元する必要があります。復元の方法には大きく2つの方法があり、出力されたリード情報のみを使う場合と、同じ生物種もしくは近縁種のゲノムDNAを参照する場合があります。前者を Assembly、後者を Reference Alignment (mapping) と呼びます。



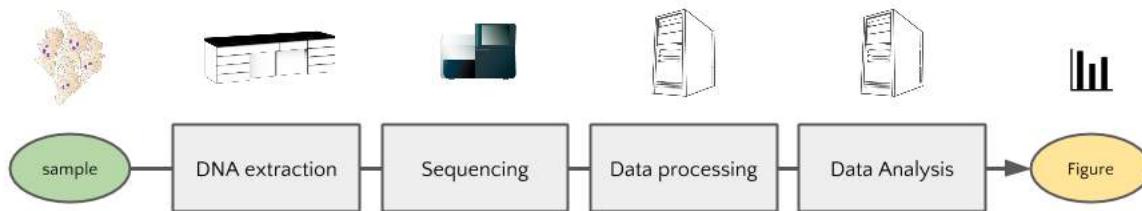
復元された塩基配列に対して、多型検出や発現量推定など、目的に応じた特徴抽出、注釈付けを行ったのち、サンプル間での統計検定を行うことで、研究データとしての解釈が可能になります。



このように、一口に「新型シーケンサーを使う」と言っても目的に応じて解釈までに様々な異なるステップを経る必要があります。目的がはっきりしていなければデータが出ても何をしていいのか分からなくなってしまいますし、逆に、目的に応じて様々な使い方をすることができる機械であるとも言えます。

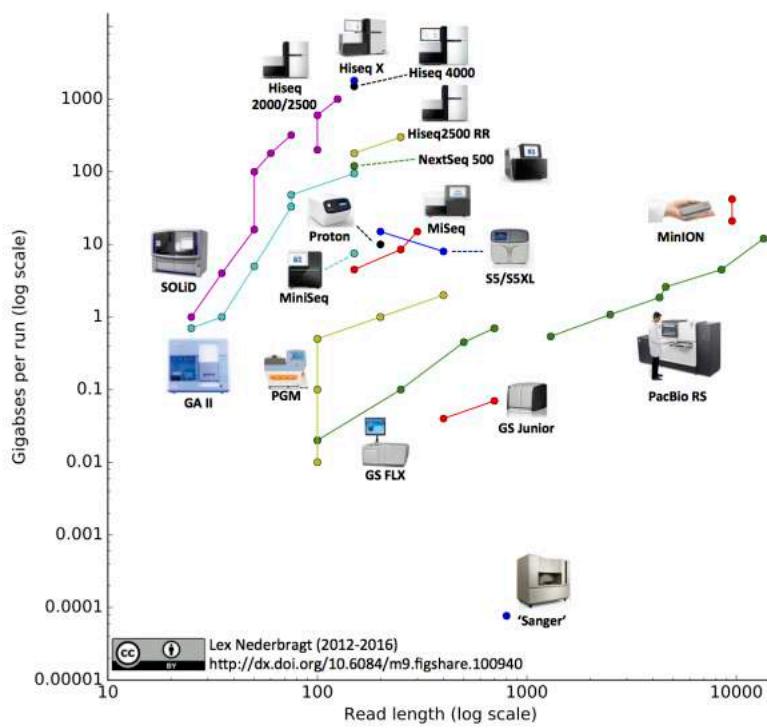


この講習では、シーケンサから出力された一次データに生物学的な注釈を行いうまでを「データ処理」、その後の統計処理などを「データ解析」と呼び分けています。人によっては両者をまとめて「データ解析」と呼ぶ場合もあります。ソフトウェアによってはデータ処理とデータ解析を一括で行うものもあります。データ処理をされる前、シーケンサーから得られた生のデータを「一次データ (primary data)」、データ処理、データ解析の過程で経られるデータを「二次データ」と呼びます。あくまでもこの講習において説明するための区別で、別の講習や書籍では異なる区分をする場合もあります。



サンプルからDNAを抽出するまではベンチワークです。単にゲノムDNAを抽出する場合もあれば、ゲノムDNAを処理して部分的に抽出したり、サンプルから抽出したRNAをDNAに逆転写したものをシーケンサーに入力として与えることもあります。詳しくは後述の「シーケンシング・アプリケーションについて」を参照してください。シーケンサーから出力されたデータから、プロットや表などの結果データを生成するまでの過程はコンピュータ上の操作が必要です。新型シーケンサーから得られるデータは一般にデータサイズが大きいため、従来のサンガーシーケンサーとは扱いが異なりますが、必要とされる計算機のスペックはサンプルの数、データのサイズ、サンプル生物種のゲノムサイズなどの要因によって異なります。

### シーケンサー機器について



新型シーケンサーから得られる塩基配列データのリードの長さや量は、機器と試薬のアップデートによって年々変化しています。また、コストあたりに得られるデータ量も変化しています。

上の図は2016年7月時点でのシーケンサー各社の公称スペックをプロットしたもので(引用: Developments in high throughput sequencing – July 2016 edition, <https://fixleblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/>)。多様なシーケンス機器が掲載されていますが、3年が経過した今、市場の競争は実質的に終結しつつあります。すなわち、短いリードでカバー率もしくはリード数を稼ぐ必要がある場合は Illumina 社のハイエンド機種を、長いリードを用いて短いリードでは測定が難しい部分の配列を決定する、あるいは新規にドラフトのゲノム配列を決定する場合には PacBio 社あるいは Oxford Nanopore 社の機種を使うことが一般的です。

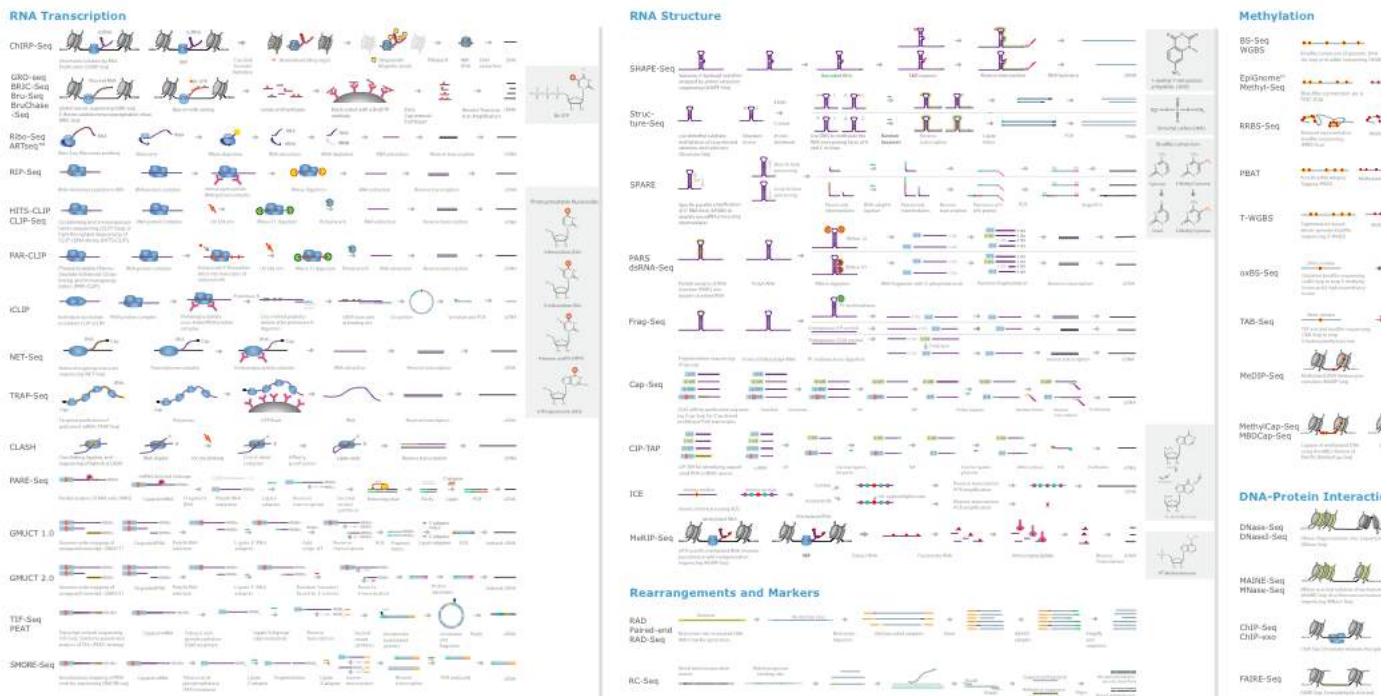
新型シーケンサーのデータを得る、もしくはデータベースに登録されたデータを利用するためには、機器の違いを理解し、目的に応じて選択する必要があります。

### シーケンシング・アプリケーションについて

DNA抽出の方法を工夫することで、新型シーケンサーの高い並列度を活かした様々な分子の計測を行うことができます。前述の通り、全ゲノムDNAを与える場合 (Whole Genome Sequencing, WGS) やエクソン領域を増幅するキットを利用した Exome シーケンス、転写物をキャプチャして逆転写したDNAをシーケンスする RNA-Seq、chromatin immunoprecipitation によって得られたDNA領域をシーケンスする ChIP-Seq などその種類は非常に多岐に渡ります。

# For all you seq...

[www.illumina.com/LibraryPrepMethods](http://www.illumina.com/LibraryPrepMethods)



アプリケーションの詳細を知るために手法についての論文を当たることが最も確実です。次の資料は Illumina 社によってまとめられたものです。

## [Sequencing Methods Review - A review of publications featuring Illumina® Technology](#)

それぞれのシーケンシング・アプリケーションにおける実験手技およびデータ処理・解析については、[次世代シーケンス解析スタンダード～NGSのポテンシャルを活かしきるWET&DRY](#)という書籍にもまとめられています。各シーケンス・アプリケーションがライラリ調製から丁寧に解説してあり、非常に参考になります。新型シーケンサーについて日本語の情報が必要な場合はまずこの一冊を購入しましょう。

本、医学・薬学・看護学・歯科学、基礎医学



## 次世代シーケンス解析スタンダード～NGSのポテンシャルを活かしきる WET&DRY 単行本 - 2014/8/23

二階堂 愛 (編集)

★★★★☆ 3件のカスタマーレビュー

その他 () の形式およびエディションを表示する

単行本

¥ 5,940

¥ 5,682 より 4 中古品の出品  
¥ 5,940 より 1 新品

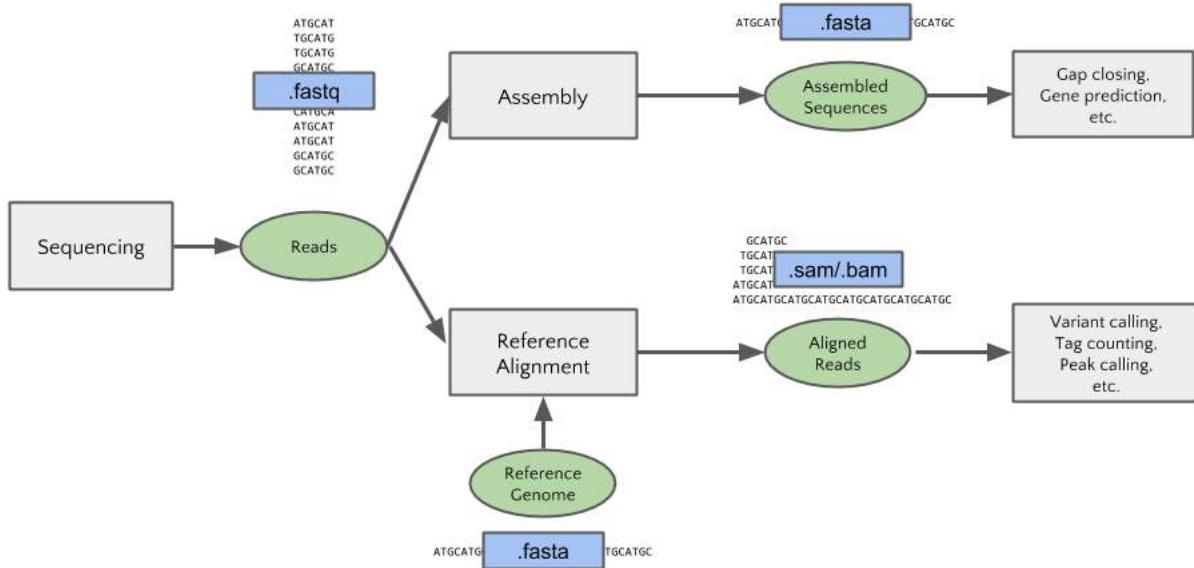
住所からお届け予定日を確認

9/1 木曜日 にお届けするには、今から 14 時間 57 分以内に「お急ぎ便」または「当日お急ぎ便」を選択して注文を確定してください (有料オプション)。Amazon プライム会員は無料)

amazon student

Amazon Student会員なら、この商品は+10%Amazonポイント還元(Amazonマーケットプレイスでのご注文は対象外)。無料体験でもれなくポイント1,000円分プレゼントキャンペーン実施中。

得られるデータについて



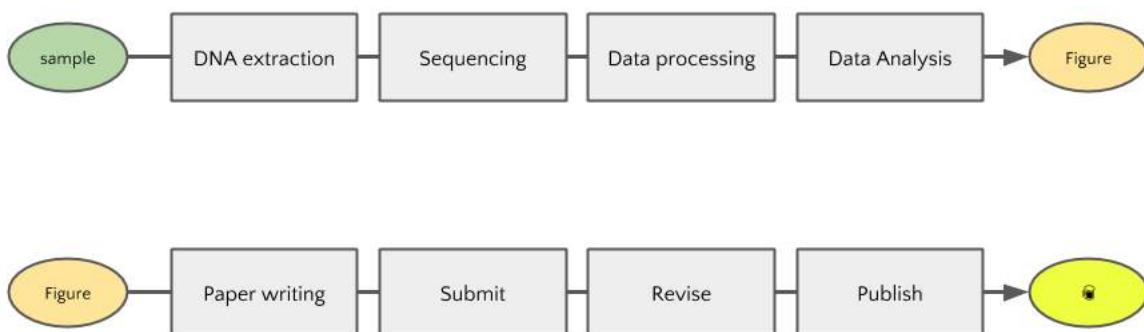
新型シーケンサのデータは一次データ(リード)から Assembly, Reference Alignmentなどのデータ処理、特徴抽出やアノテーションなどのデータ解析を経る中で、様々なファイルフォーマットで保存されます。ソフトウェアごとにどのフォーマットを入力/出力するかが大抵の場合決まっているので、フォーマットを理解しておくことが重要です。

フォーマットには色々な種類があるため、手に入れたデータをどのように見ていいか分からず場合は、オンラインで公開されているファイルフォーマットの仕様を参照してください。ゲノムブラウザでお馴染みのUCSCに、データフォーマットについてまとめられたページがあり便利です。<https://genome.ucsc.edu/FAQ/FAQformat.html>

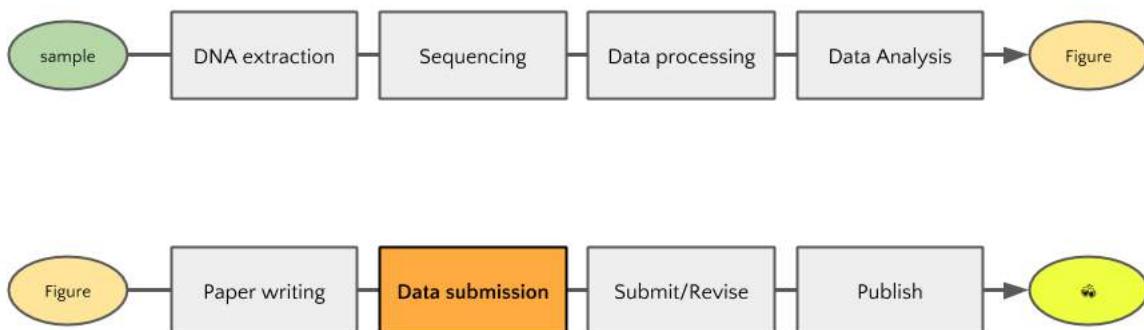
## 新型シーケンサーのデータベースとは

### データが登録されるまで

シーケンサーから得られたデータはデータ処理、データ解析を経て可視化され、可視化されたデータは論文あるいは何かしらのアウトプットの一部となることが一般的です。



しかし、現在ではほとんどの論文誌において、論文に用いた一次データ(つまりシーケンサから得られたデータ処理前のデータ)は投稿前に公共データレポジトリに登録してアクセス用のリンクを取得する必要があります。



配列データを公共データレポジトリのサーバに送信し、サンプルやシーケンス実験についてのメタデータをガイドラインに従って記述する必要があります。つまり、発表された論文で使われたデータは(理想的には)その全てがオンラインで公開されていることになります。

配列データの登録は思っているよりも大変です。日本の登録窓口である [DDBJ](#) では専門のアノテーターが登録作業を補助してくれますが、数日かかることもあります。論文投稿前ではなく、データ解析や処理をする前の余裕のある時に登録しておきましょう。登録後はすぐに公開されるわけではなく、任意の期間(原則2年間) プライベートにしておくことができます。

## 一次データレポジトリ

新型シーケンサーから得られたデータは個人を特定可能な情報を含むヒトデータとそれ以外に分けられ、データレポジトリに登録されます。

### Open Access Data

個人特定可能な情報を含むヒトデータ以外のデータは全て Sequence Read Archive (SRA) への登録を指定されます。SRA は International Nucleotide Sequence Database Collaboration (INSDC) によって運用されています。INSDCに参加しているのは米国の National Center for Biotechnology Information (NCBI), 欧州の European Bioinformatics Institute (EBI), 日本の DNA Data Bank of Japan (DDBJ) の3つの機関です。

3つの機関はそれぞれが独自にデータの登録・検索のシステムを公開・運用していますが、登録されたデータは3つの機関で交換され共有されています。

The screenshot shows the NCBI SRA homepage. At the top, there's a navigation bar with 'NCBI Resources' and 'How To'. Below it is a search bar with 'SRA' selected and an 'Advanced' link. The main content area has a large orange background with white sequence data (e.g., 'GG TTAAG ATACATAAATTT AATAC AACGCC TTGCATTAG TAA CGCCCT') and a dark blue sidebar with the word 'SRA'. The sidebar text explains the purpose of SRA: 'Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.' Below this are three columns: 'Getting Started' with links like 'How to Submit', 'Login to SRA', etc.; 'Tools and Software' with links like 'Download SRA Toolkit', 'SRA Toolkit Documentation', etc.; and 'Related Resources' with links like 'Submission Portal', 'Trace Archive', etc.

[NCBI SRA - www.ncbi.nlm.nih.gov/sra](http://NCBI SRA - www.ncbi.nlm.nih.gov/sra)

EMBL-EBI

# ENA

European Nucleotide Archive

Services | Research | Training

Search Examples: BN000065, histone

Advanced Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support | Feedback

## European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#)

Access to ENA data is provided through the browser, through search tools, large scale file download and through the API.

### Text Search

Examples: BN000065, histone

[Search](#)

[Advanced search](#)

### Sequence Search

Enter or paste a nucleotide sequence or accession number

[Search](#)

[Advanced search](#)

### Popular

- o [Submit and update](#)
- o [Sequence submissions](#)
- o [Genome assembly submissions](#)
- o [Submitting environmental sequences](#)
- o [Citing ENA data](#)
- o [Rest URLs for data retrieval](#)
- o [Rest URLs to search ENA](#)

### Latest ENA news

**03 Aug 2016:** Projects and studies merged in the ENA browser

Projects and studies have now been merged within the ENA browser so that there is a single landing page.

**02 Aug 2016:** Scheduled disruption to ENA services

Due to planned electrical maintenance work at EBI between 26th and 30th August, there will be wide-spread disruptions to ENA services.

**22 Jun 2016:** ENA Release 128

Release 128 of ENA's assembled/annotated sequence data is now available.

[EBI ENA - ebi.ac.uk/ena](#)

DDBJ

## Sequence Read Archive

Login & Submit | Databases | English | Contact | Google 検索

Home | Handbook | FAQ | Search | Download | Pipeline | About DRA

DDBJ Sequence Read Archive (DRA) は Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD® Systemなどの次世代シーケンサからの出力データのためのデータベースです。DRA は International Nucleotide Sequence Database Collaboration (INSDC) のメンバーであり、NCBI Sequence Read Archive (SRA) と EBI Sequence Read Archive (ERA) との国際協力のもと、運営されています。従来のキャビラリ式シーケンサからの出力データは DDBJ Trace Archive にご登録ください。

[検索](#) [登録](#)

[DDBJ DRA - trace.ddbj.nig.ac.jp/dra](#)

### データの登録の仕方と登録されたデータの見方

SRAにデータを登録するためには、データがどのようにレポジトリ内で整理され保管されているかを知る必要があります。データを探す場合にも、必要なデータを探すためにはどのようにデータのメタデータが記述されているかを知ることが重要です。以下が登録の手順を簡略化したものです。

1. 登録アカウントを作成してデータファイルをアップロードする
2. BioProject にプロジェクトの情報作成・登録する
3. BioSample にサンプルのレコードを作成・登録する
4. SRAにメタデータを登録する
  1. Experiment のレコードを登録する
  2. Run のレコードを登録する
5. データの検証をパスすれば登録完了、アクセション番号が発行される

配列データはファイルとして SRA のサーバで(つまり NCBI, EBI, DDBJ いずれにも)公開されており、ダウンロードすることができますが、その配列データがどのようなサンプルから得られたか、どのようなシーケンスをしたか、という情報、つまりメタデータは細かい単位に分けられ、それぞれにIDを振って管理されています。つまり、以下のようになります。

- プロジェクトの情報
  - ID: PRJDA38027
    - <http://www.ncbi.nlm.nih.gov/bioproject/PRJDA38027>
- サンプルの情報
  - ID: SAMD00016353
    - <http://www.ncbi.nlm.nih.gov/biosample/SAMD00016353>
- 実験の情報

- ID: DRX000001
  - <http://trace.ddbj.nig.ac.jp/DRASearch/experiment?acc=DRX000001>
- Run の情報
  - ID: DRR000001
    - <http://trace.ddbj.nig.ac.jp/DRASearch/run?acc=DRR000001>

メタデータがこのように細かく分かれている理由は、個別のサンプルや個別の実験をレポート側で管理しやすくするためですが、登録する際やデータを検索する際にはちょっと複雑です。複雑ですがめげずにやっていきましょう。

### Controlled Access Data

配列データやメタデータにサンプルを提供した個人が特定可能な情報が含まれている場合には、データアクセスに許可が必要なデータ、すなわち controlled access データとなります。アクセスするための許可を得る方法は、登録されているデータレポート、およびデータセットごとに異なります。また、fastq データではなく変異情報のサマリーデータのみが登録される場合もあります。

The screenshot shows the dbGaP (Database of Genotypes and Phenotypes) homepage. At the top, there's a navigation bar with links for NCBI, Resources, How To, and a search bar. Below that is a header with the dbGaP logo and a search dropdown. Underneath is a large image of a human eye and brain. The main content area has three columns: 'Access dbGaP Data' (with links to Advanced Search, Controlled Access Data, Public FTP Download, Collections, and Summary Statistics), 'Resources' (with links to Phenotype-Genotype Integrator, Association Results Browser, dbGaP RSS Feed, Software, and dbGaP Tutorial), and 'Important Links' (with links to How to Submit, FAQ, Code of Conduct, Security Procedures, and Contact Us). Below these sections is a 'Latest Studies' table:

Study	Embargo Release	Details	Participants	Type Of Study	Links	P
<a href="#">phs001074.v1.p1</a> GeneSTAR NextGen Functional Genomics of Platelet Aggregation	Version 1: passed embargo		250	Longitudinal Cohort, Family	<a href="#">Links</a>	<a href="#">Human1Mv1_C</a>
<a href="#">phs000937.v1.p1</a> The Landscape of Antisense Gene Expression in Human Cancers	Version 1: passed embargo		376	Cohort	<a href="#">Links</a>	<a href="#">HiSeq 2000</a>
<a href="#">phs001151.v1.p1</a> Region-specific Transcriptome Analysis of the Human Retina and RPE/Choroid	Version 1: passed embargo		5	Control Set	<a href="#">Links</a>	<a href="#">TruSeq Strand+ Kit</a> <a href="#">HiSeq</a>

[NCBI dbGaP - www.ncbi.nlm.nih.gov/gap](#)

# European Genome-phenome Archive

All

Examples: EGAS00000000001, Cancer

[EGA home](#) | [About](#) | [Studies](#) | [Datasets](#) | [Data access committees](#) | [Data providers](#) | [Submit to EGA](#) | [Contact Us](#)

The European Genome-phenome Archive (EGA) allows you to explore **datasets** from genomic **studies**, provided by a range of **data providers**. Access to datasets must be approved by the specified **Data Access Committee (DAC)**.

[Nature Genetics 47, 692–695, \(2015\) | doi:10.1038/ng.3312](#)

## Help

- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)

**ADVANCE NOTICE:** Due to planned **electrical maintenance work at EBI 26th - 30th August**, there will be wide-spread disruptions to EGA services. All submission services will be unavailable, including submission upload dropboxes. The EGA Helpdesk ([ega-helpdesk@ebi.ac.uk](mailto:ega-helpdesk@ebi.ac.uk)) will also be unavailable. Download services are not likely to be affected. **Normal service should be resumed by the 31st August.**

## Studies

Studies are experimental investigations of a particular phenomenon or trait.

[Browse all studies](#)

## Learn about the EGA

- [Introduction to the EGA](#)
- [How to obtain an account with the EGA](#)
- [Using your EGA account](#)

## Datasets

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).

[Browse all datasets](#)

[Browse all control datasets](#)

## Navigation

- [Login](#)
- [Request new password](#)



## Data Access Committees

Providers may be involved in study creation, submission and designation of Data Access Committees (DACs).

EBI EGA - [www.ebi.ac.uk/ega/](http://www.ebi.ac.uk/ega/)

**NBDC** NBDCヒトデータベース

ホーム データの利用 データの提供 ガイドライン NBDCヒトデータ審査委員会

NBDCヒトデータベースについて

ヒトに関するデータは、次世代シーケンサーはじめとした解析技術の発達に伴って膨大な量が产生されつつあり、それらを整理・格納して、生命科学の進展のために有効に活用するためのルールや仕組みが必要です。

国立研究開発法人科学技術振興機構(JST)バイオサイエンスデータベースセンター(NBDC)では、個人情報の保護に配慮しつつヒトに関するデータの共有や利用を推進するために、ヒトに関する様々なデータを共有するためのプラットフォーム『NBDCヒトデータベース』を設立するとともに、[国立遺伝学研究所 DNA Data Bank of Japan \(DDBJ\)](#)と協力して、ヒトに関するデータを公開しています。

本Webサイトを通じて、ヒトに関するデータの利用及びヒトに関するデータの提供を行なうことができます。データ共有についての概要は[こちら](#)を参照下さい。

新着情報

2016/08/26 群馬大学大学院 医学系研究科 かほく（）を公開しました (hum0064)

2016/08/01 九州大学病院 別府病院 外科 かほく（） (hum0026)

2016/07/28 大阪大学大学院 歯学研究科 かほく（） (hum0027)

▶ ニュース一覧へ

Search NBDC Human Database Beacon for Alternative Alleles [API help](#)

A NBDC Human Database Beacon will be a member of [GA4GH Beacon Network](#).

GRCh37 e.g. 12:112241766 A Search Example: ALDH2 Variant (GRCh37, 12:112241766 A)

## 利用可能な研究データ一覧

データ利用方法は[こちら](#)をご覧下さい。

Research ID	研究題目	公開日	データの種類	研究方法	手法	参考 (対象)
hum001.v1 JGAS00000000002	SCA31罹患患者のゲノム解析データ	v1:2013/12/01	NGS (WGS)	配列決定	Illumina (HiSeq 2000)	SCA31：1症例 (日本人)
hum003.v1 DRA000908	関節リウマチ患者及び健常人におけるHLA領域の塩基配列比較解析	v1:2013/07/01	NGS (Target Capture)	HLA領域 配列決定	Illumina (MiSeq)	33検体 (セルライン)
hum004.v1 JGAS00000000001	上皮成長因子受容体遺伝子異変を有する肺腺癌の細胞生物学的遺伝子変異プロファイル	v1:2014/07/11	NGS (Exome)	配列決定	Illumina (GAIIx)	肺腺癌：97症例 (日本人)

[NCBI JGA - humandb.biosciencedbc.jp](#)

### Controlled Access Data にはどのようにアクセスするのか

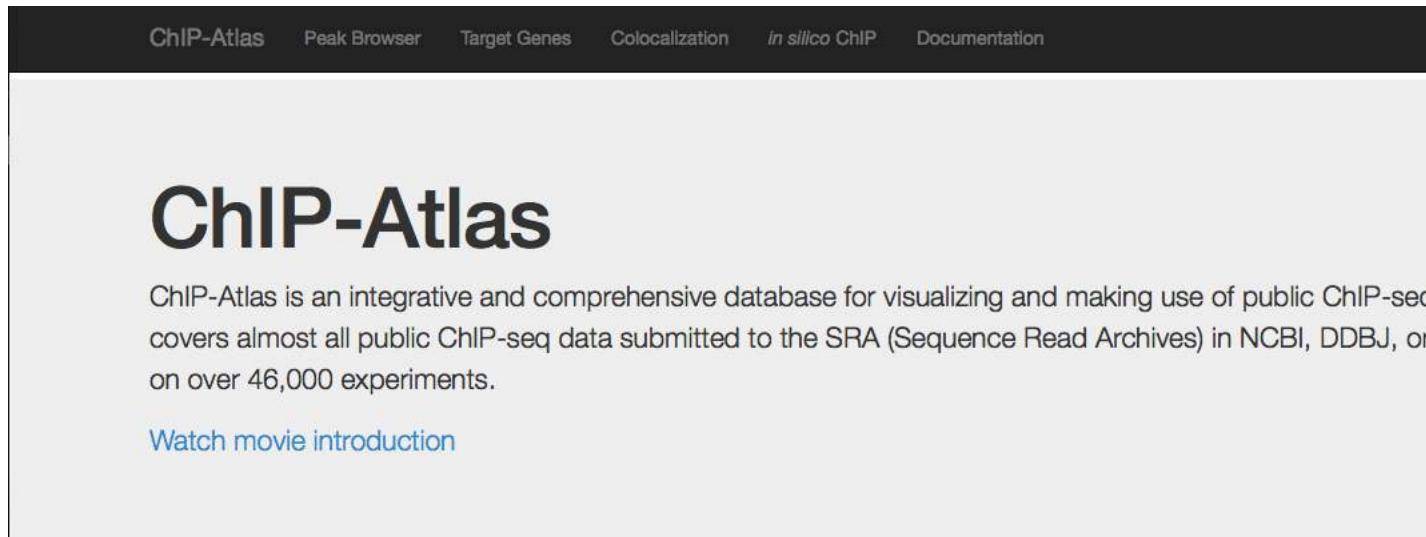
データの検索はオンラインで可能ですが、実際のデータをダウンロードしたりそれを元に研究する場合には許可を得る必要があります。どこから許可を得るかは、レポジトリごとに異なり、以下のようにになります。

- dbGaP の場合
  - National Institute of Health (NIH) の審査を受ける必要があります。
    - <https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>
- EGA の場合
  - 利用したいデータごとに、データを寄託した組織・団体による data access-granting organisation (DAO) からの審査を受ける必要があります。
    - <https://www.ebi.ac.uk/ega/about>
- JGA の場合
  - NBDCヒトデータ審査委員会 の審査を受ける必要があります。
    - <http://humandbs.biosciencedbc.jp/data-use>

### 二次データを含むデータベース

SRAに登録されたデータを再利用したい場合には、データをダウンロードしたのち、データ処理・データ解析を行う必要があります。データ処理に用いられるソフトウェアの利用法を解説する書籍やウェブサイトも増え(後述します)、データ処理や解析の敷居は低くなっています。しかし、SRAに登録されたデータを予め処理したものや、それに基づく解析の結果を公開するデータベースも増えています。また、国際コンソーシアムなどによって行われる大型プロジェクトでは、プロジェクトのウェブサイトを通じて解析済みデータを公開している場合があります。

#### ChIP-Atlas



The ChIP-Atlas homepage features a dark header with the following navigation links: ChIP-Atlas, Peak Browser, Target Genes, Colocalization, *In silico* ChIP, and Documentation. Below the header, a large, bold title 'ChIP-Atlas' is centered. A brief description follows: 'ChIP-Atlas is an integrative and comprehensive database for visualizing and making use of public ChIP-seq data. It covers almost all public ChIP-seq data submitted to the SRA (Sequence Read Archives) in NCBI, DDBJ, or EMBL-EBI, on over 46,000 experiments.' Below the description is a blue link 'Watch movie introduction'.

The four main features of ChIP-Atlas are:

#### Peak Browser

graphically visualizes protein binding on given genomic loci with genome browser (IGV).

[Watch Movie](#)

#### Target Genes

predicts target genes bound by given transcription factors.

[Watch Movie](#)

#### Colocalization

predicts partner proteins colocalizing with given transcription factors.

[Watch Movie](#)

#### *In silico* ChIP

predicts protein loci and gene

[Watch Movie](#)

[ChIP-Atlas](#) は 九州大医学部 発生再生研究室 と DBCLS によって開発され、九州大によって運用されているデータベースです。SRAで公開された ChIP-Seq, DNase-Seq のデータに定型処理を行い、その結果データに基いて様々な解析結果を公開しています。

ChIP-Atlas Peak Browser Target Genes Colocalization *in silico* ChIP Documentation

## ChIP-Atlas - Peak Browser

Visualize All Peaks from Published ChIP-Seq data.

H. sapiens M. musculus D. melanogaster C. elegans S. cerevisiae

Antigen Class		Cell type Class		Threshold for Significance
All antigens (19281)	DNase-seq (1074)	Muscle (197)	Neural (543)	50
Histone (4962)	RNA polymerase (777)	Pancreas (208)	Placenta (21)	100
TFs and others (5914)	Input control (2361)	Pluripotent stem cell (2081)	Prostate (839)	200
Unclassified (817)	No description (3376)	Uterus (856)	Others (433)	500

Antigen		Cell type		View
Pou5F1	POU5F1	type to search	Embryonic Stem Cells (21)	View
BACH1 (1)	ES-3 (1)	NCCIT (15)	NT2-D1 (3)	Download
BCL11A (2)	hESC BG03 (3)	hESC Cyt49 (1)	hESC H1 (116)	
BRCA1 (1)	hESC			
BRD2 (2)				
BRD3 (2)				

ChIP-Atlas Peak Browser Target Genes Colocalization *in silico* ChIP Documentation Find a

## ChIP-Atlas - *in silico* ChIP

Analyze your data with public ChIP-seq data.

H. sapiens M. musculus D. melanogaster C. elegans S. cerevisiae

1. Antigen Class		2. Cell type Class		3. Threshold for Significance
All antigens (19281)	DNase-seq (1074)	All cell types (19281)	Adipocyte (194)	50
Histone (4962)	RNA polymerase (777)	Blood (5892)	Bone (243)	100
TFs and others (5914)	Input control (2361)	Breast (2082)	Breast (2082)	200
Unclassified (817)	No description (3376)	Cardiovascular (585)	Digestive tract (1329)	500

4. Select your data		5. Select dataset to be compared		6. Describe datasets
<input checked="" type="radio"/> Genomic regions (BED) or sequence motif <small>①</small>	<input type="radio"/> Gene list (Gene symbols) <small>①</small>	<input checked="" type="radio"/> Random permutation of user data <small>①</small>	Permutation times <input type="radio"/> x1 <input type="radio"/> x10 <input type="radio"/> x100	User data title <small>①</small>
chr8 134339107 134339325	chr15 41548749 41548861	<input type="radio"/> BED or sequence motif <small>①</small>		My data
chr15 25207262 25207383	chr1 190561786 190561937			Compared data title <small>①</small>
chr3 142208484 142208615	chr9 115544616 115544762			Control
chr15 59846350 59846560	chr15 81109164 81109289			Project title <small>①</small>
<small>[Choose File]</small> no file selected	Choose local file			My project
				<b>submit</b>
				Estimated run time: 5 mins

# wPGSA

Estimate relative activities of transcriptional regulators from transcriptome data by weighted Parametric Analysis: wPGSA.

## Upload your transcriptome data to run wPGSA

It takes a few minutes or longer depend on your data size.

[See result page of](#)

No file selected

## About input data format

Log fold change data should be prepared in an ASCII tab delimited text file. It is organized as follows.

To create and edit the log fold change file, use a text editor or Excel. When you use Excel, be aware of a problem as described in [Zeeberg et al 2003](#).



[wPGSA online](#) は 理化学研究所 医科学イノベーションハブ推進プログラム 川上英良博士と DBCLS のコラボレーションで開発・運用しているウェブサービスです。遺伝子発現データを入力として与えると、与えられた発現データに関与している転写因子の予測を行います。この手法では、SRAに登録されたChIP-Seqデータの再解析結果が利用されています。

## MicrobeDB.jp

[MicrobeDB.jp](#) は統合データベースプロジェクト・微生物統合データベースによって開発・運用されているデータベースです。SRAで公開されている メタ16S、メタゲノム、メタトランスクriプトームなどのデータの解析データが蓄積されており、ブラウザ上で可視化することができます。

Gene: psbA  
 Taxonomy: Streptococcus glycerinaceus  
 Mapping: Escherichia coli O157:H7 str. Sakai  
 Environment: hot spring  
 SRS: rumen  
 Strain: Bifidobacterium  
 Disease: Cholera  
 MiGap: GAF

 Microbe DB<sup>JP</sup>
 
[Sign In](#)

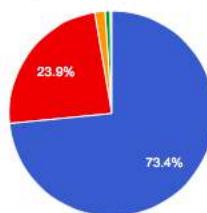
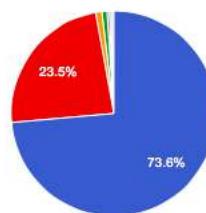
[Definition](#) [SRS008091](#) [help](#)

### Metadata

Body habitat	gut
Investigation type	16S rRNA
Project name	The Microbiota Mediates Pathogen Clearance from the Gut Lumen after Non-Typhoidal Salmonella Diarrhea
Sample title	no title
Scientific name	mouse gut metagenome
Sequencing method	pyrosequencing
Target gene	16S rRNA V5-V6 regions
Taxon id	<a href="#">410661</a>

### Taxonomic Composition of Meta16S

Domain
Phylum
Class

EMBL-EBI

**About us News**

Overview Leadership Funding Background Collaborations Jobs People & groups News Events Visit us Contact us

[About us](#) > [News](#) > [Service news](#) > RESTful RNA-seq Analysis API version 1.2 is out

## RESTful RNA-seq Analysis API version 1.2 is out



Gene & Protein  
Expression

Are you interested in annotating a genome from high-quality RNA-seq data, performing a meta-analysis of gene expression in human tissues, or simply wonder what RNA-seq data is available in the public domain? It's all there for you, just one API call away.

The newly released version 1.2 of the RESTful RNA-seq Analysis API serves ftp locations of CRAM, bigWig and bedGraph files, as well as raw, FPKM and TPM gene and exon counts for 265,000 public sequencing runs in 264 species.

This release includes analysis results for 38 new entries from WormBase ParaSite, as well as aggregated baseline gene expression in tissues, cell types, developmental stages, sex and strains in 61 organisms.

Give our RESTful RNA-seq Analysis API a try! You can find the [RESTful RNA-seq Analysis documentation](#) on our website.

### News

- News overview
- News archive
- Brochure
- Photos & images
- Blogs

<p>EMBL-EBI</p> <p>News Brochures Contact us Intranet</p>	<p><b>Services</b></p> <p>By topic By name (A-Z) Help &amp; Support</p>	<p><b>Research</b></p> <p>Overview Publications Research groups Postdocs &amp; PhDs</p>	<p><b>Training</b></p> <p>Overview Train at EBI Train outside EBI Train online Contact organisers</p>	<p><b>Industry</b></p> <p>Overview Members Area Workshops SME Forum Contact Industry programme</p>
---	---	---	---	--

EBI RNA-Seq Analysis API <http://www.ebi.ac.uk/fg/rnaseq/api/> は、SRAに登録されたもののうち、264生物種、265,000のシーケンスデータを解析した結果データを取得することができるサービスです。bigWig, bedGraph, FPKM, TPMなどの情報を取ることができます。

API とは "Application Programming Interface" の略で、コンピューター・プログラムからアクセスされることを意図してデザインされたシステムであることを意味します。そのため、通常のウェブサイトのような綺麗な見た目や検索機能などはありません。しかし、APIのアクセスパターンを理解すれば、プログラムが書けなくてもウェブブラウザでデータを取得することができます。さらに、Shell Script や Ruby, Python などの比較的簡単なプログラミング言語を覚えれば、より効率よく処理を行うことができます。(興味のある人はトライしてみてください。たのしいよ! )

EBI RNA-Seq Analysis API の仕様書は <http://www.ebi.ac.uk/fg/rnaseq/api/doc> にあります。EBI RNA-Seq Analysis API は EBI の遺伝子発現データベースである Expression Atlas の一部として開発されています。論文は以下を参照してください。

Petryszak, Robert, et al. "Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants." *Nucleic acids research* (2015): gkv1045.

EBI RNA-Seq Analysis API は以下の4つの種類のデータを取得することができます。

- Analysis Results Per Run
- Analysis Results Per Study
- Sample Attributes Per Run
- Baseline Expression Per Gene - for Tissue, Cell Type, Developmental Stage, Sex and Strain

それぞれ、アクセスするURLのパターンによって欲しいデータを取得することができます。例として、"Baseline Expression Per Gene" 機能で遺伝子ごとのデータを取得してみましょう。以下のURLにウェブブラウザでアクセスしてみてください。

[http://www.ebi.ac.uk/fg/rnaseq/api/tsv/50/getExpression/homo\\_sapiens/POU5F1](http://www.ebi.ac.uk/fg/rnaseq/api/tsv/50/getExpression/homo_sapiens/POU5F1)

“

Mac/Linuxなら「ターミナル」を開いて、  
curl "[http://www.ebi.ac.uk/fg/rnaseq/api/tsv/50/getExpression/homo\\_sapiens/POU5F1](http://www.ebi.ac.uk/fg/rnaseq/api/tsv/50/getExpression/homo_sapiens/POU5F1)"  
と打ち込んでみてください。

GENE_ID	ORGANISM	MEDIAN_EXPRESSION		COEFFICIENT_OF_VARIATION		NUMBER_OF_RUNS	ORGANISM
ENSG00000204531	homo_sapiens	160.6	0.5	190	embryo	NA	embryonic day 4
ENSG00000204531	homo_sapiens	145.70000000000002		0.5	377	embryo	NA
ENSG00000204531	homo_sapiens	39.80000000000004		0.6	235	pooled tumor	NA
ENSG00000204531	homo_sapiens	30.6	1.7	415	embryo	NA	embryonic day 6
ENSG00000204531	homo_sapiens	9.9	2.3000000000000003		466	embryo	NA
ENSG00000204531	homo_sapiens	8.9	0.8	81	embryo	NA	embryonic day 3
ENSG00000204531	homo_sapiens	7.8	1.3	168	soft tissue	NA	NA
ENSG00000204531	homo_sapiens	7.3	1.1	130	cortex (temporal lobe)	neuron	NA
ENSG00000204531	homo_sapiens	5.3	0.9	97	lung	NA	NA
ENSG00000204531	homo_sapiens	5.1000000000000005		1.8	50	pancreas	NA
ENSG00000204531	homo_sapiens	4.6000000000000005		0.7000000000000001		60	lung
ENSG00000204531	homo_sapiens	3.7	0.9	102	lung	NA	male
ENSG00000204531	homo_sapiens	3.3000000000000003		0.5	89	skin of lower leg	NA
ENSG00000204531	homo_sapiens	3.2	0.6	55	skin of lower leg	NA	female
ENSG00000204531	homo_sapiens	3	0.4	249	blood	NA	female
ENSG00000204531	homo_sapiens	2.8000000000000003		1.4000000000000001		92	liver
ENSG00000204531	homo_sapiens	2.7	0.3	112	blood	NA	male
ENSG00000204531	homo_sapiens	2.4	0.3	81	tibial nerve	NA	male
ENSG00000204531	homo_sapiens	2.3000000000000003		0.4	55	thyroid	NA
ENSG00000204531	homo_sapiens	2.2	2.3000000000000003		80	testis	NA
ENSG00000204531	homo_sapiens	2.2	0.8	95	tibial artery	NA	male
ENSG00000204531	homo_sapiens	2.1	0.8	58	aorta	NA	male
ENSG00000204531	homo_sapiens	2.1	0.6	92	subcutaneous adipose tissue	NA	NA
ENSG00000204531	homo_sapiens	2.1	0.4	81	thyroid	NA	male
ENSG00000204531	homo_sapiens	2	0.5	52	subcutaneous adipose tissue	NA	NA
ENSG00000204531	homo_sapiens	2	1.2	119	blood	NA	NA
ENSG00000204531	homo_sapiens	2	0.9	140	prostate	NA	NA
ENSG00000204531	homo_sapiens	2	0.5	54	tibial artery	NA	female
ENSG00000204531	homo_sapiens	1.9000000000000001		1.2	71	mucosa of esophagus	NA
ENSG00000204531	homo_sapiens	1.8	1.5	175	microdissected cortical-like ventricle from cere		
http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByCondition/1548							
ENSG00000204531	homo_sapiens	1.8	4.7	333	dissociated whole cerebral organoid	NA	1
ENSG00000204531	homo_sapiens	1.6	0.8	478	whole blood	NA	NA
ENSG00000204531	homo_sapiens	1.6	0.5	565	bone marrow	acute myeloid leukemia	NA
ENSG00000204531	homo_sapiens	1.5	0.8	52	respiratory airway	airway basal cell	1
ENSG00000204531	homo_sapiens	1.5	1	69	pancreatic islet	NA	male
ENSG00000204531	homo_sapiens	1.5	0.8	58	blood	granulocyte	NA
ENSG00000204531	homo_sapiens	1.3	0.8	608	heparinised blood	acute myeloid leukemia	1
ENSG00000204531	homo_sapiens	1.3	2.1	226	fetal neocortex	NA	NA
ENSG00000204531	homo_sapiens	1.3	1.5	246	brain	NA	NA
ENSG00000204531	homo_sapiens	1.3	0.6	52	lymph node	NA	adult
ENSG00000204531	homo_sapiens	1.3	0.8	119	skin	NA	NA
ENSG00000204531	homo_sapiens	1.2	0.7000000000000001		80	placental villus	parenchyma
ENSG00000204531	homo_sapiens	1.2	1	162	whole blood	NA	male
ENSG00000204531	homo_sapiens	1.2	1	77	whole blood	NA	female
ENSG00000204531	homo_sapiens	1.1	1.3	59	breast	NA	NA
ENSG00000204531	homo_sapiens	1	0.5	68	esophagus muscularis mucosa	NA	NA
ENSG00000204531	homo_sapiens	1	0.4	103	transformed fibroblast	NA	male
ENSG00000204531	homo_sapiens	0.9	0.6	77	heart left ventricle	NA	male
ENSG00000204531	homo_sapiens	0.8	0.5	57	transformed fibroblast	NA	female
ENSG00000204531	homo_sapiens	0.7000000000000001		0.6	73	brain (ba9 prefrontal cortex)	1
ENSG00000204531	homo_sapiens	0.7000000000000001		0.6	69	ba9 prefrontal cortex	NA
ENSG00000204531	homo_sapiens	0.6	0.8	183	skeletal muscle	NA	male
ENSG00000204531	homo_sapiens	0.6	0.9	141	skeletal muscle	NA	female
ENSG00000204531	homo_sapiens	0.5	0.9	52	adipose	NA	NA
ENSG00000204531	homo_sapiens	0.5	1.5	55	na	NA	NA
ENSG00000204531	homo_sapiens	0.3	1.3	80	left ventricle apex	tissue	NA
ENSG00000204531	homo_sapiens	0.1	2.1	289	blood	Thrombocytes	NA

こんな結果が表示されます。画面に表示されたデータは tsv (tab separated values) です。これをそのままコピー・アンド・ペーストで Google Spreadsheet なんかに貼り付けて見るのがよいでしょう。自分のパソコンに保存したい場合は、画面上で右クリックして「名前を付けて保存」をするとよいです。この場合は "ebirnaseq.expression.50.homo\_sapiens.POU5F1.tsv" なんて名前を付けてみてはどうでしょう。

このAPIでは取得したいデータの条件をURLの中に埋め込んでいきます。このURLに埋め込まれた条件は次の通りです。

- <http://>
  - ハイパーテキストransfer protocolです。
- [www.ebi.ac.uk/fg/rnaseq/api](http://www.ebi.ac.uk/fg/rnaseq/api)
  - EBI RNA-Seq Analysis API のサービスの base URL です。
- /tsv
  - フォーマットを tsv に指定しています。json にもできます。
- /50

◦ minimum number of runs の値です。APIは条件に合致したエントリ(プロジェクト)を返しますが、エントリごとにRunの数が違います。この値は、表示するエントリをRun数で絞り込むことができます。50を指定することで「50回以上シーケンスRunのあるエントリを表示」という条件になります。

- /getExpression
  - 4つの機能のうちの Expression Per Gene のデータを取得することを示します。
- /homo\_sapiens
  - 生物種を指定しています。
- /POU5F1
  - 遺伝子名を gene symbol で指定しています。

上のURLをちょっと変更することで別のデータを取得することができます。色々なデータを取得して比べてみたり、別のデータと組み合わせてみたりしてみてください。

- json 形式にしてみる
  - [http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/homo\\_sapiens/POU5F1](http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/homo_sapiens/POU5F1)
- Run数での絞り込みをやめてみる
  - [http://www.ebi.ac.uk/fg/rnaseq/api/json/0/getExpression/homo\\_sapiens/POU5F1](http://www.ebi.ac.uk/fg/rnaseq/api/json/0/getExpression/homo_sapiens/POU5F1)
- 別の遺伝子にしてみる
  - [http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/homo\\_sapiens/SOX2](http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/homo_sapiens/SOX2)
- 別の生物種にしてみる
  - [http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/mus\\_musculus/pou5f1](http://www.ebi.ac.uk/fg/rnaseq/api/json/50/getExpression/mus_musculus/pou5f1)

表示されたデータの最後のカラムには "ALL\_SAMPLE\_ATTRIBUTES" というカラム名でURLが各行に埋め込まれています。これはサンプルの情報を取得するためのAPIのURLです。1つ選んでアクセスしてみましょう。

GENE_ID	ORGANISM	MEDIAN_EXPRESSION	COEFFICIENT_OF_VARIATION	NUMBER_OF_RUNS	ORGANISM
ENSG00000204531	homo_sapiens	160.6	0.5	190	embryo NA embryonic day 4 NA NA
ENSG00000204531	homo_sapiens	145.70000000000002	0.5	377	embryo NA embryonic day 5 NA
ENSG00000204531	homo_sapiens	39.80000000000004	0.6	235	pooled tumor NA NA
ENSG00000204531	homo_sapiens	30.6	1.7	415	embryo NA embryonic day 6 NA NA
ENSG00000204531	homo_sapiens	9.9	2.3000000000000003	466	embryo NA embryonic day 7 NA
ENSG00000204531	homo_sapiens	8.9	0.8	81	embryo NA embryonic day 3 NA NA
ENSG00000204531	homo_sapiens	7.8	1.3	168	soft tissue NA NA NA
ENSG00000204531	homo_sapiens	7.3	1.1	130	cortex (temporal lobe) neuron NA NA
ENSG00000204531	homo_sapiens	5.3	0.9	97	lung NA NA NA
ENSG00000204531	homo_sapiens	5.1000000000000005	1.8	50	pancreas NA NA
ENSG00000204531	homo_sapiens	4.6000000000000005	0.7000000000000001	60	lung NA
ENSG00000204531	homo_sapiens	3.7	0.9	102	lung NA male NA
ENSG00000204531	homo_sapiens	3.3000000000000003	0.5	89	skin of lower leg NA NA
ENSG00000204531	homo_sapiens	3.2	0.6	55	skin of lower leg NA NA
ENSG00000204531	homo_sapiens	3	0.4	249	blood NA female NA
ENSG00000204531	homo_sapiens	2.8000000000000003	1.4000000000000001	92	liver NA
ENSG00000204531	homo_sapiens	2.7	0.3	112	blood NA male NA
ENSG00000204531	homo_sapiens	2.4	0.3	81	tibial nerve NA male NA
ENSG00000204531	homo_sapiens	2.3000000000000003	0.4	55	thyroid NA female NA
ENSG00000204531	homo_sapiens	2.2	2.3000000000000003	80	testis NA male NA
ENSG00000204531	homo_sapiens	2.2	0.8	95	tibial artery NA male NA
ENSG00000204531	homo_sapiens	2.1	0.8	58	aorta NA male NA
ENSG00000204531	homo_sapiens	2.1	0.6	92	subcutaneous adipose tissue NA NA
ENSG00000204531	homo_sapiens	2.1	0.4	81	thyroid NA male NA
ENSG00000204531	homo_sapiens	2	0.5	52	subcutaneous adipose tissue NA NA
ENSG00000204531	homo_sapiens	2	1.2	119	blood NA NA NA
ENSG00000204531	homo_sapiens	2	0.9	140	prostate NA NA NA
ENSG00000204531	homo_sapiens	2	0.5	54	tibial artery NA female NA
ENSG00000204531	homo_sapiens	1.9000000000000001	1.2	71	mucosa of esophagus NA
ENSG00000204531	homo_sapiens	1.8	1.5	175	microdissected cortical-like ventricle from cere
<a href="http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByCondition/1548">http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByCondition/1548</a>					
ENSG00000204531	homo_sapiens	1.8	4.7	333	dissociated whole cerebral organoid NA
ENSG00000204531	homo_sapiens	1.6	0.8	478	whole blood NA NA NA
ENSG00000204531	homo_sapiens	1.6	0.5	565	bone marrow acute myeloid leukemia NA
ENSG00000204531	homo_sapiens	1.5	0.8	52	respiratory airway airway basal cell
ENSG00000204531	homo_sapiens	1.5	1	69	pancreatic islet NA NA male
ENSG00000204531	homo_sapiens	1.5	0.8	58	blood granulocyte NA NA NA
ENSG00000204531	homo_sapiens	1.3	0.8	608	heparinised blood acute myeloid leukemia
ENSG00000204531	homo_sapiens	1.3	2.1	226	fetal neocortex NA NA NA
ENSG00000204531	homo_sapiens	1.3	1.5	246	brain NA NA NA
ENSG00000204531	homo_sapiens	1.3	0.6	52	lymph node adult NA NA
ENSG00000204531	homo_sapiens	1.3	0.8	119	skin NA NA NA
ENSG00000204531	homo_sapiens	1.2	0.7000000000000001	80	placental villus parenchyma
ENSG00000204531	homo_sapiens	1.2	1	162	whole blood NA NA male NA
ENSG00000204531	homo_sapiens	1.2	1	77	whole blood NA NA female NA
ENSG00000204531	homo_sapiens	1.1	1.3	59	breast NA NA NA
ENSG00000204531	homo_sapiens	1	0.5	68	esophagus muscularis mucosa NA NA
ENSG00000204531	homo_sapiens	1	0.4	103	transformed fibroblast NA NA male
ENSG00000204531	homo_sapiens	0.9	0.6	77	heart left ventricle NA NA male
ENSG00000204531	homo_sapiens	0.8	0.5	57	transformed fibroblast NA NA female
ENSG00000204531	homo_sapiens	0.7000000000000001	0.6	73	brain (ba9 prefrontal cortex)
ENSG00000204531	homo_sapiens	0.7000000000000001	0.6	69	ba9 prefrontal cortex NA
ENSG00000204531	homo_sapiens	0.6	0.8	183	skeletal muscle NA NA male NA
ENSG00000204531	homo_sapiens	0.6	0.9	141	skeletal muscle NA NA female NA
ENSG00000204531	homo_sapiens	0.5	0.9	52	adipose NA NA NA
ENSG00000204531	homo_sapiens	0.5	1.5	55	na NA NA NA
ENSG00000204531	homo_sapiens	0.3	1.3	80	left ventricle apex tissue NA NA
ENSG00000204531	homo_sapiens	0.1	2.1	289	blood Thrombocytes NA NA

<http://www.ebi.ac.uk/fg/rnaseq/api/tsv/getSampleAttributesByCondition/1373>

STUDY_ID	RUN_ID	TYPE	VALUE	EFO_URL
ERP012552	ERR1041424	cell	22	NA
ERP012552	ERR1041424	developmental stage	embryonic day 4	NA
ERP012552	ERR1041424	individual	E4.1	NA
ERP012552	ERR1041424	inferred lineage	NA	NA
ERP012552	ERR1041424	inferred pseudo-time	8.4769277	NA
ERP012552	ERR1041424	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041424	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041424	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922
ERP012552	ERR1041424	phenotype	NA	NA
ERP012552	ERR1041424	single cell well quality	OK	NA
ERP012552	ERR1041424	treatment	No	NA
ERP012552	ERR1041425	cell	23	NA
ERP012552	ERR1041425	developmental stage	embryonic day 4	NA
ERP012552	ERR1041425	individual	E4.1	NA
ERP012552	ERR1041425	inferred lineage	NA	NA
ERP012552	ERR1041425	inferred pseudo-time	8.338401525	NA
ERP012552	ERR1041425	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041425	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041425	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922
ERP012552	ERR1041425	phenotype	NA	NA
ERP012552	ERR1041425	single cell well quality	OK	NA
ERP012552	ERR1041425	treatment	No	NA
ERP012552	ERR1041426	cell	24	NA
ERP012552	ERR1041426	developmental stage	embryonic day 4	NA
ERP012552	ERR1041426	individual	E4.1	NA
ERP012552	ERR1041426	inferred lineage	NA	NA
ERP012552	ERR1041426	inferred pseudo-time	8.34096982	NA
ERP012552	ERR1041426	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041426	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041426	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922
ERP012552	ERR1041426	phenotype	NA	NA
ERP012552	ERR1041426	single cell well quality	OK	NA
ERP012552	ERR1041426	treatment	No	NA
ERP012552	ERR1041427	cell	25	NA
ERP012552	ERR1041427	developmental stage	embryonic day 4	NA
ERP012552	ERR1041427	individual	E4.1	NA
ERP012552	ERR1041427	inferred lineage	NA	NA
ERP012552	ERR1041427	inferred pseudo-time	6.578250692	NA
ERP012552	ERR1041427	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041427	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041427	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922
ERP012552	ERR1041427	phenotype	NA	NA
ERP012552	ERR1041427	single cell well quality	OK	NA
ERP012552	ERR1041427	treatment	No	NA
ERP012552	ERR1041428	cell	26	NA
ERP012552	ERR1041428	developmental stage	embryonic day 4	NA
ERP012552	ERR1041428	individual	E4.1	NA
ERP012552	ERR1041428	inferred lineage	NA	NA
ERP012552	ERR1041428	inferred pseudo-time	9.626029352	NA
ERP012552	ERR1041428	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041428	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041428	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922
ERP012552	ERR1041428	phenotype	NA	NA
ERP012552	ERR1041428	single cell well quality	OK	NA
ERP012552	ERR1041428	treatment	No	NA
ERP012552	ERR1041429	cell	27	NA
ERP012552	ERR1041429	developmental stage	embryonic day 4	NA
ERP012552	ERR1041429	individual	E4.10	NA
ERP012552	ERR1041429	inferred lineage	NA	NA
ERP012552	ERR1041429	inferred pseudo-time	6.625735271	NA
ERP012552	ERR1041429	inferred trophectoderm subpopulation	NA	NA
ERP012552	ERR1041429	organism	Homo sapiens	http://purl.obolibrary.org/obo/NCBITaxon
ERP012552	ERR1041429	organism part	embryo	http://purl.obolibrary.org/obo/UBERON_0000922

表示されたのはエントリに含まれるシーケンスRunとそのサンプルの情報です。STUDY ID と Run ID は SRA ID ですね。Run ID を使って、発現量のカウントを実行した後のデータが取れないかやってみましょう。"Analysis Results Per Run" を使います。以下のURLにアクセスしてみてください。

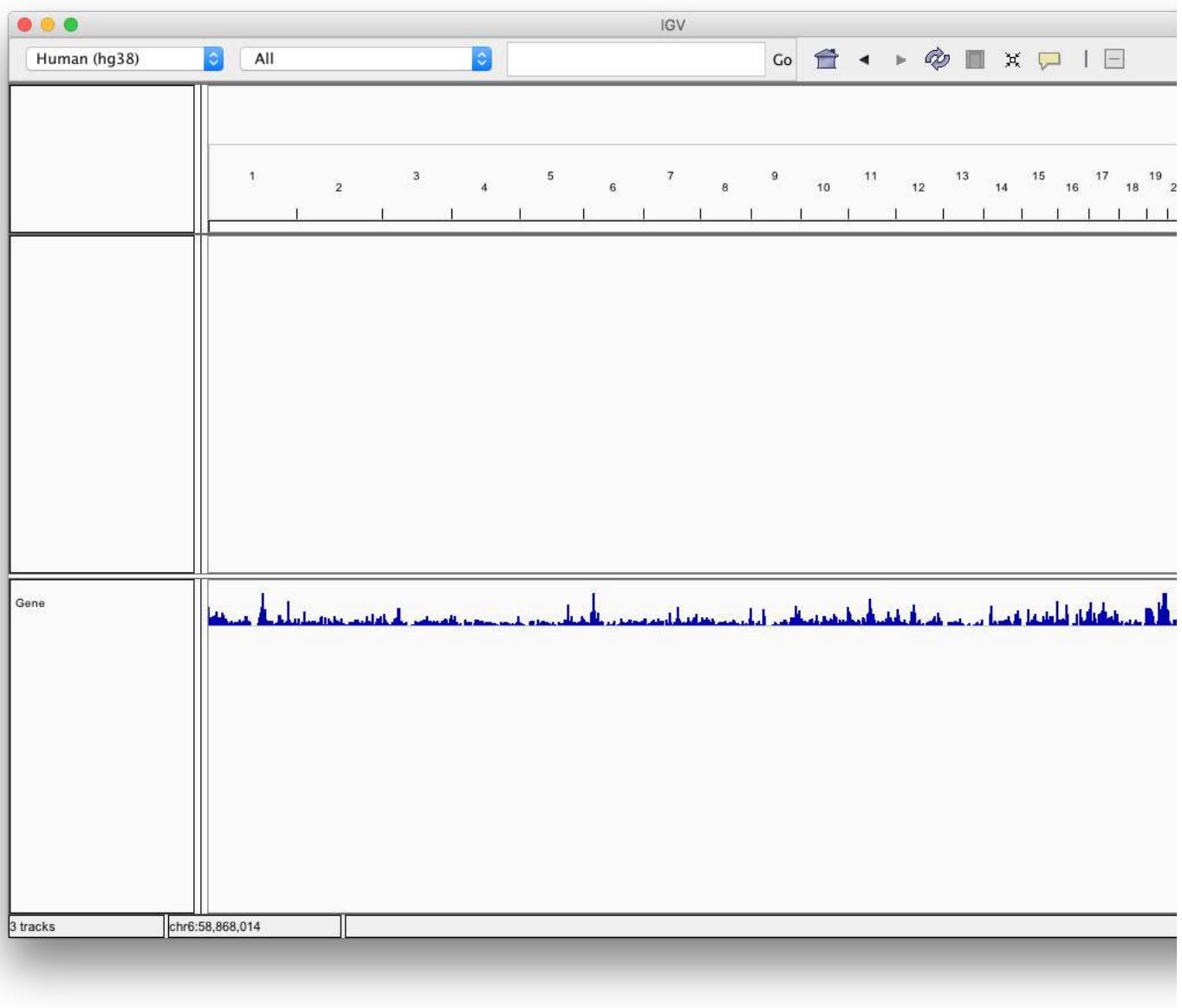
<http://www.ebi.ac.uk/fg/rnaseq/api/tsv/70/getRun/ERR1041424>

STUDY_ID	SAMPLE_IDS	BIOREP_ID	RUN_IDS	ORGANISM	REFERENCE_ORGANISM	STATUS
BIGWIG_LOCATION	MAPPING_QUALITY					
ERP012552	SAMEA3580795	ERR1041424	ERR1041424	homo_sapiens	homo_sapiens	Complete
ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/data/atlas/rnaseq/ERR104/004/ERR1041424/ERR1041424.cram						
ftp://ftp.ebi.ac.uk/pub/databases/arrayexpress/data/atlas/rnaseq/ERR104/004/ERR1041424/ERR1041424.bw						

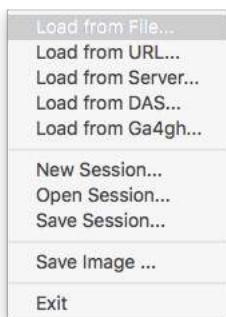
いくつかURLが表示されますが、".bw"で終わるftpのURLをコピーしましょう。".bw"はbigWigファイルであることを示す拡張子です(bigWigフォーマットについては[こちら](#))。ここでこのファイルを可視化するためにゲノムブラウザである[Integrative Genomics Viewer](#)を使ってみます(デモ時のバージョンは2.3.81です)。

先ほどのgetRunの結果のページのデータ中に、ゲノムのバージョンはGRCh38であると書いてあるので、IGVでもゲノムはGRCh38を選択します。ダウンロードしたてのIGVにはhg18とhg19しか含まれていないので、初回だけゲノムデータをダウンロードする必要があります。

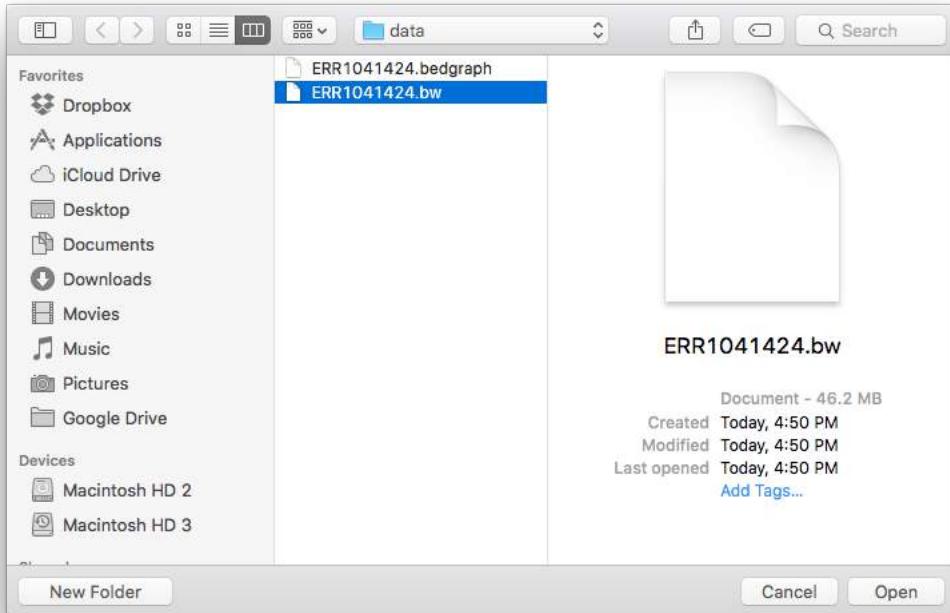
IGVではURLを入力するとデータを取得し、表示してくれる機能がありますが、今回のデータはサイズが大きくタイムアウトエラー(ファイルの取得に時間がかかりすぎる)が出て怒られるので、bigWigデータを手元にダウンロードします。ブラウザでURLにアクセスするだけでデータのダウンロードが始まります。データのダウンロードが完了したら、IGVのメニューからFile > Load from File...を選択し、ダウンロードしたbigWigファイルを選択します。



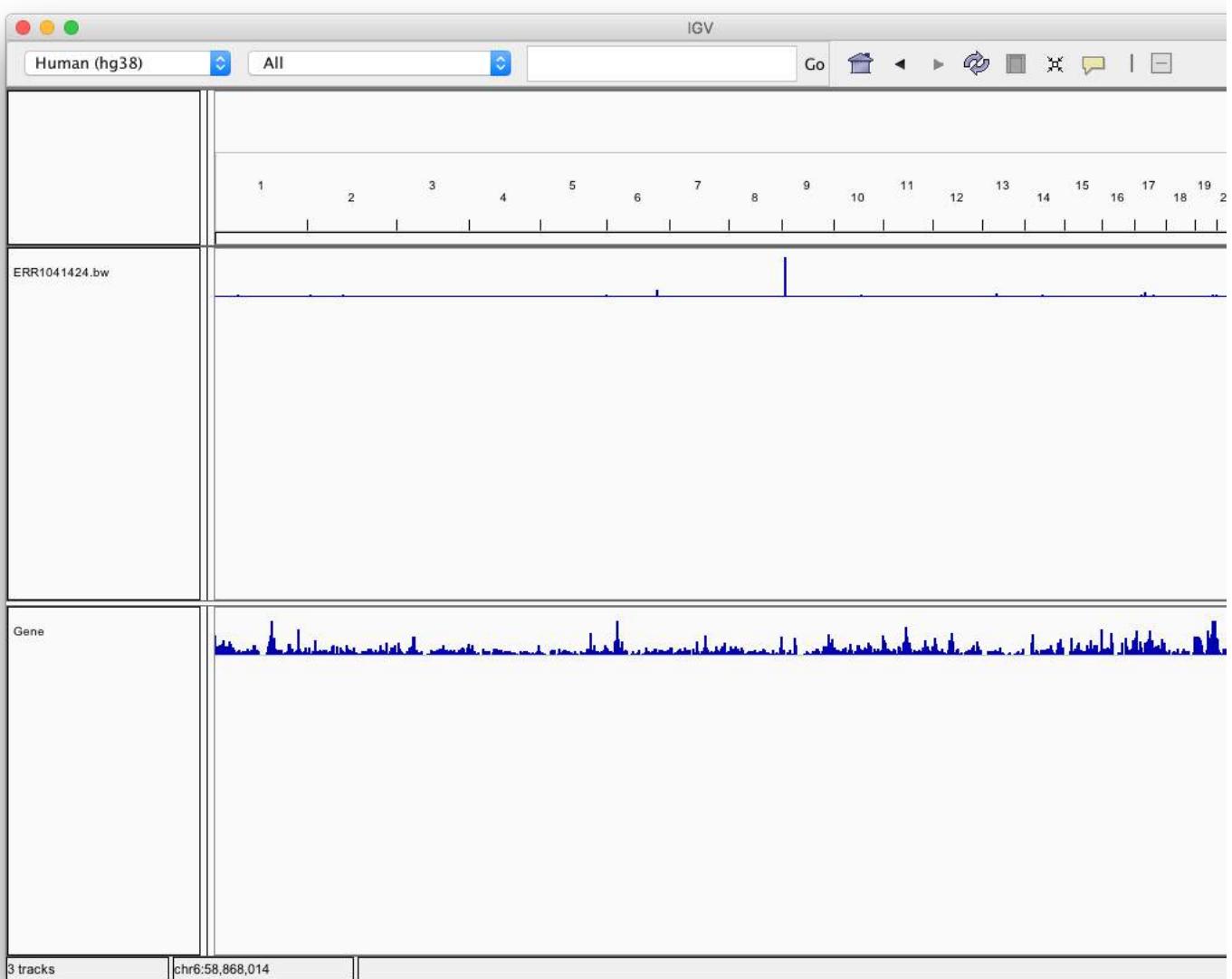
IGV を開いて GRCh38 を選択したところです。



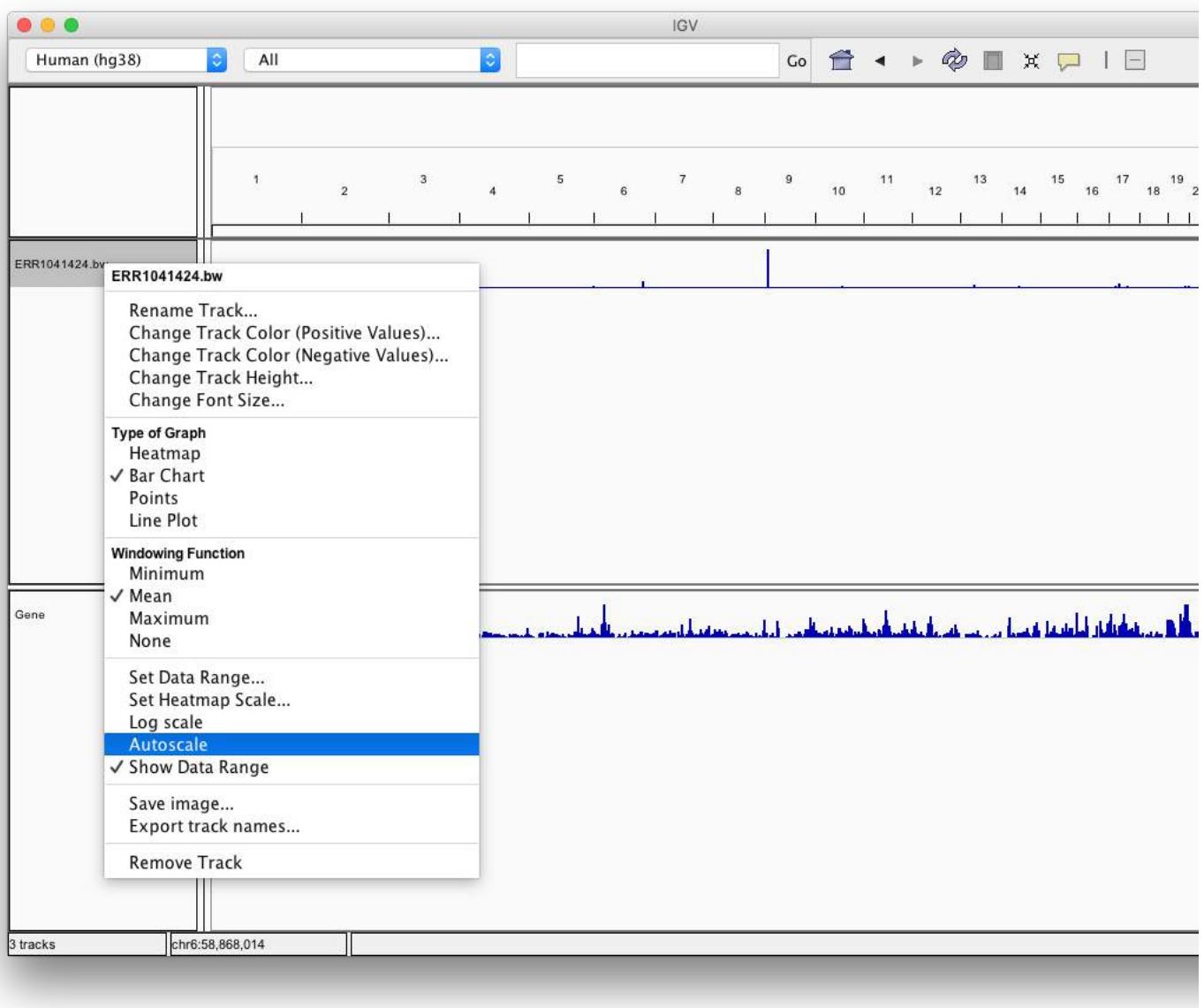
メニューから File > Load from File... を選んで



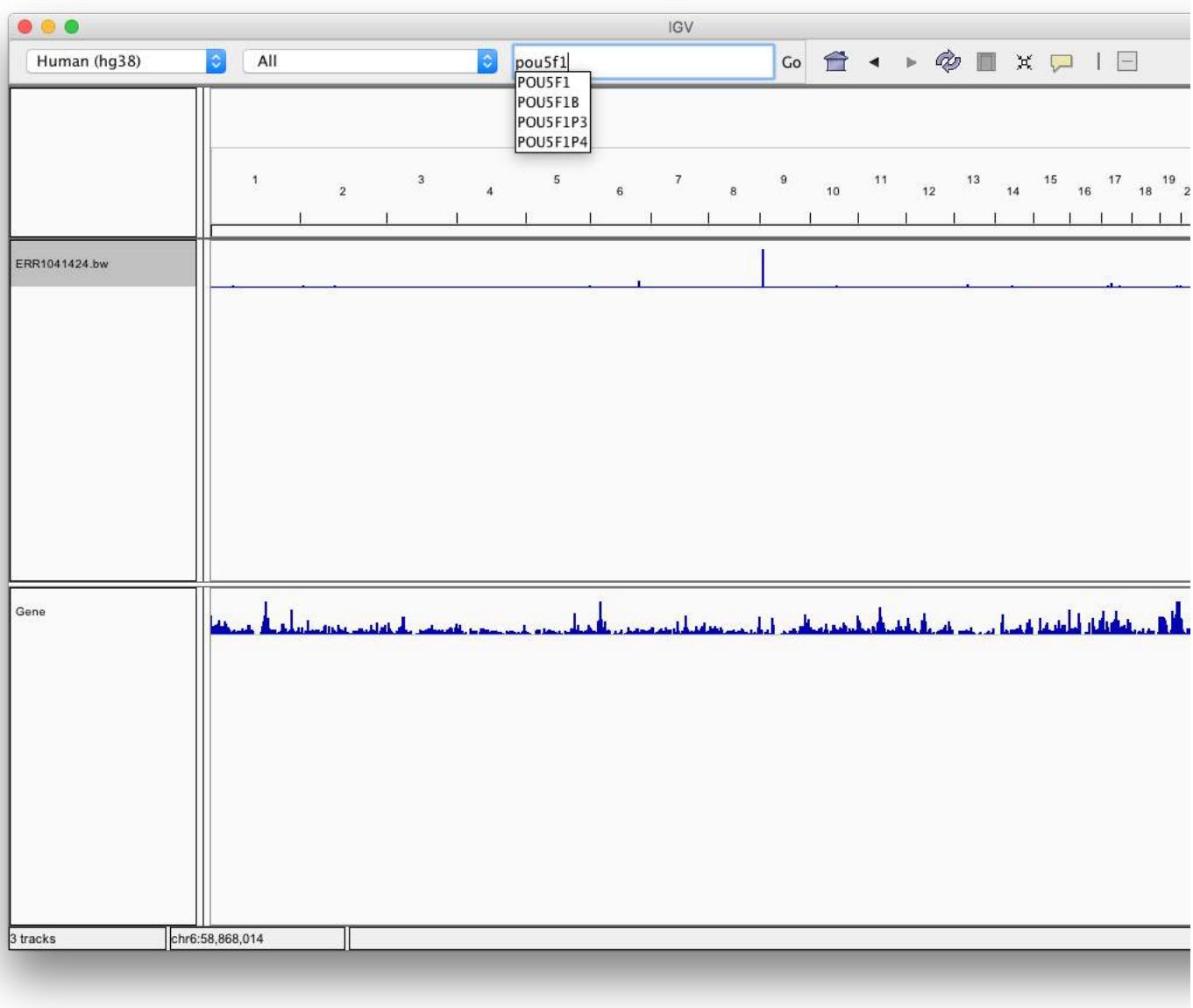
ダウンロードした .bw ファイルを選びます。



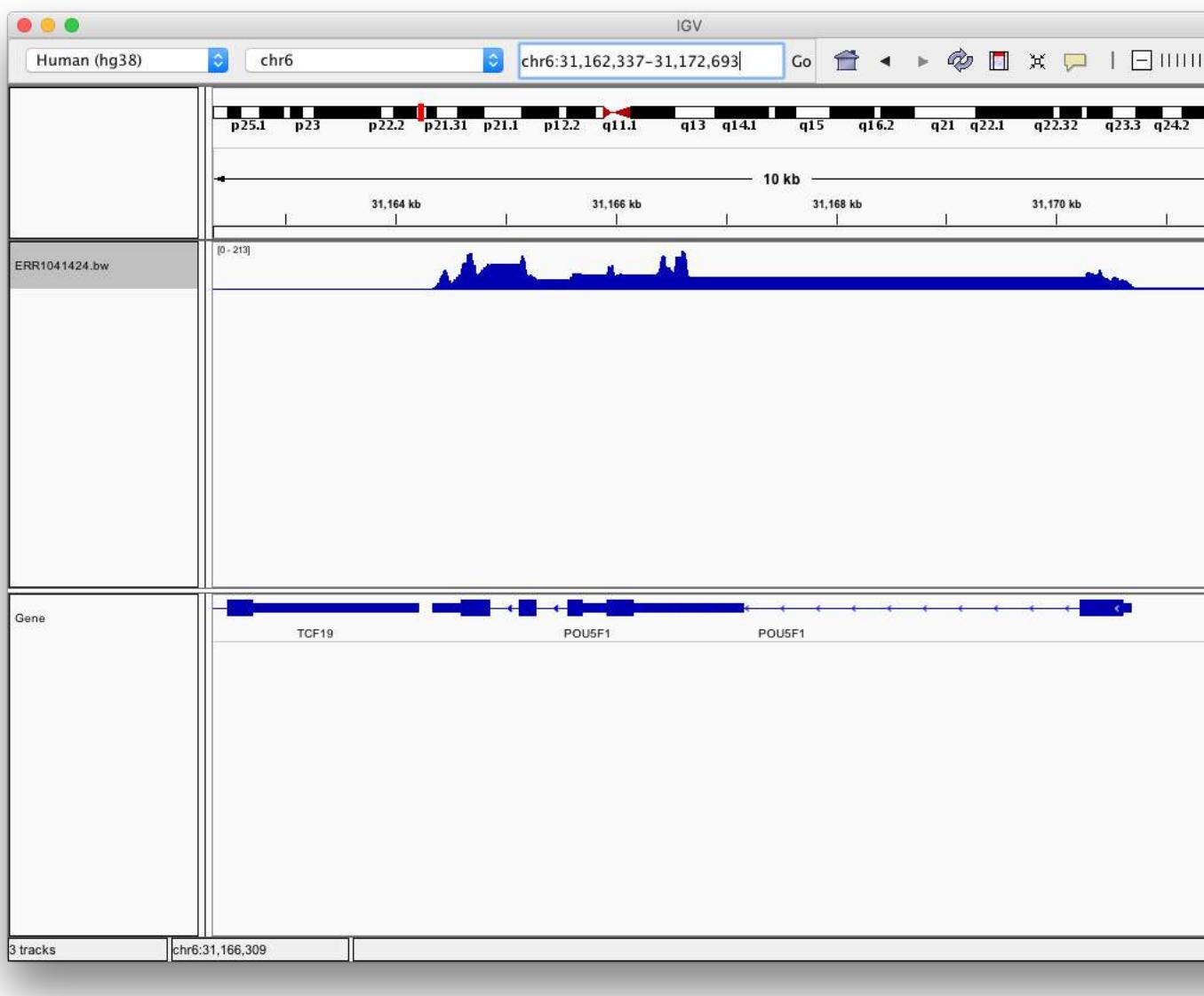
読み込まれました。



読み込まれたトラックで右クリックを押して "Auto Scale" を選択しておきます。



POU5F1 の様子を見てみましょう。検索ボックスに POU5F1 と入力します。



遺伝子発現のデータが表示されました！

このように、SRAに登録されている発現データを自分で解析しなくても、APIを使ってアクセスし、データをダウンロードすることで、手元で可視化したり、他のデータと比較したりすることができます。

## データ解析についてのTips

解析済みのデータの共有が進み簡単にデータを二次利用することができるようになっています。しかし、データによっては解析データが存在しなかったり、特定のソフトウェアの結果が必要な場合もあります。その場合は、SRAなどの一次データレポジトリからデータをダウンロードして自分で解析する必要があります。

コンピュータを用いたデータ解析に不慣れな方向けに、多くの教材や書籍、講習会などがあります。自分のデータを自分で解析してみたいという場合には、以下を参考にしてみてください。

- 統合データベース講習会
  - 講習資料と講演動画が [統合TV - togotv.dbcls.jp](#) から公開されています。
- NGSハンズオン講習会
  - <http://biosciencedbc.jp/human/human-resources/workshop/h28-2>
- DDBJが提供するオンラインの解析サービスの使い方
  - [Pipeline Tutorial](#)
  - [統合TV](#)
    - [今日からはじめるDDBJ Read Annotation Pipeline](#)
    - [DDBJ Read Annotation Pipelineによるde novo Assembly解析](#)
    - [DDBJパイプラインとGalaxyによるデータ解析](#)
- 書籍: [次世代シーケンサーDRY解析教本\(細胞工学別冊\)](#)

## 以上で終了です！

おつかれさまでした！