

Transforming genomic coordinates between plant reference assemblies

www.gramene.org

plants.ensembl.org

Kapeel Chougule, MS
Cold Spring Harbor Laboratory
December 6, 2016



Assembly Conversion Tool

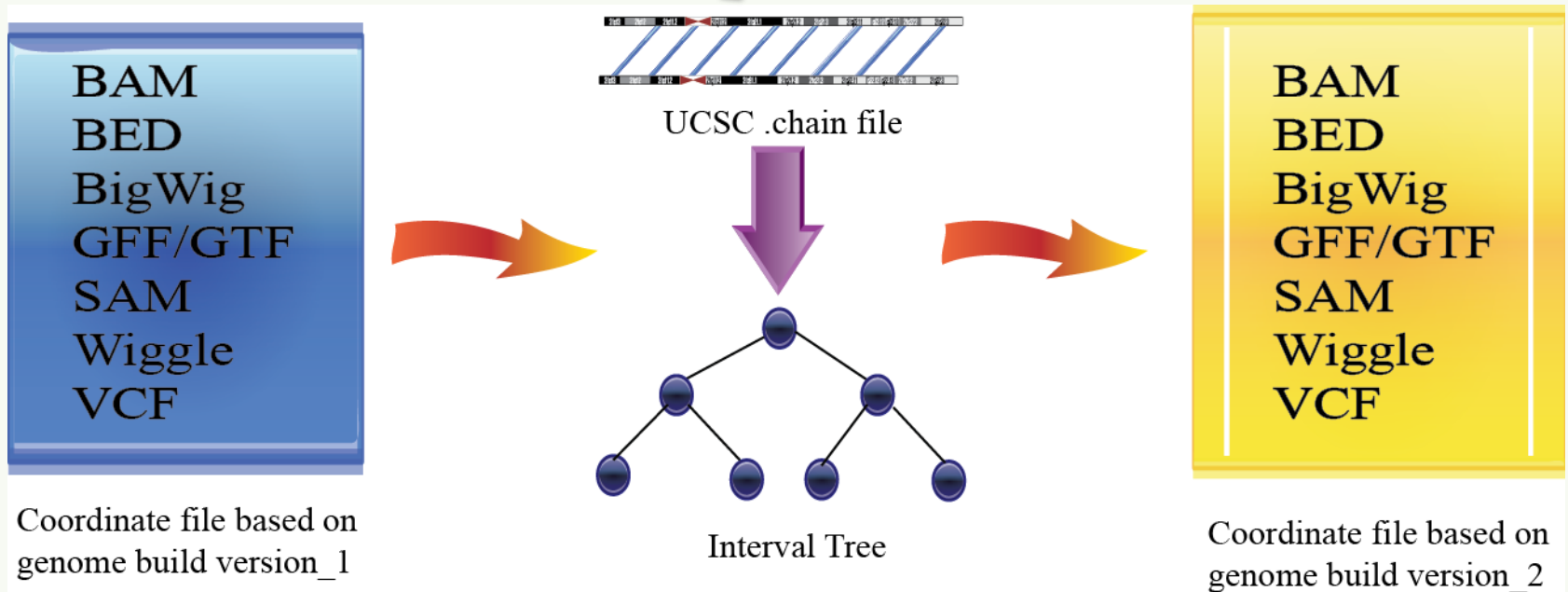
Converts genome coordinates between assemblies to facilitate meta-analysis, direct comparison, data integration and visualization

Why?

- Reference genome assemblies are subjected to change from time to time
- Convert results from old assemblies to newer versions and *viceversa*

CrossMap Tool

Assembly converter tool in Gramene uses CrossMap!!



- Chain file describes the pair-wise alignments between two genomes
- CrossMap run time increases linearly to the size of input file

Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., & Wang, L. (2013). CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* (Oxford, England), btt730

Other conversion tools

- [UCSC liftover tool](#) only supports BED input
- [NCBI remap](#) support BED, GFF, GTF, VCF, etc.
- [Galaxy](#) (based on UCSC liftover tool) supports BED, GFF, GTF input
- [Pyliftover](#) only does conversion of point coordinates; unlike liftOver, it does not convert ranges, nor does it provide any special facilities to work with BED files

Tools: Assembly converter







[BLAST](#)[BioMart](#)[Tools](#)[Downloads](#)[Help](#)[Feedback](#)[UploadData](#)

Tools

We provide a number of ready-made tools for processing both our data and yours. We routinely delete results from our servers after 10 days, but if you have an [gramene genome account](#) you will be able to save the results for about 2-3 months until next release.





Processing your data

<http://ensembl.gramene.org/tools.html>

Name	Description	Online tool	Upload limit	Download script	Documentation
Variant Effect Predictor 	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.		50MB*		
HMMER	Quickly search our genomes for your protein sequence.				
BLAST/BLAT	Search our genomes for your DNA or protein sequence.		50MB		
Assembly Converter	Map (liftover) your data's coordinates to the current assembly.		50MB		

* For larger datasets we provide an API script that can be downloaded (you will also need to install our Perl API, below, to run the script).

Accessing Gramene data

Name	Description	Get it from:	Documentation
BioMart	Use this data-mining tool to export custom datasets from Gramene.	Gramene BioMart	
Ensembl Perl API	Programmatic access to all Ensembl data using simple Perl scripts	GitHub or FTP download (current release only)	
Ensembl Genomes Virtual Machine	Pre-configured VirtualBox virtual machine (VM) running the latest Ensembl Genomes browser.		
Ensembl Genomes REST server	Access Ensembl data using your favourite programming language		



Tools: Assembly converter

This online tool currently uses [CrossMap](#), which supports a limited number of formats (see our online documentation for [details of the individual data formats](#) listed below). CrossMap also discards metadata in files, so track definitions, etc, will be lost on conversion.

Important note: CrossMap converts WIG files to BedGraph internally for efficiency, and also outputs them in BedGraph format.

Species:

Assembly mapping:

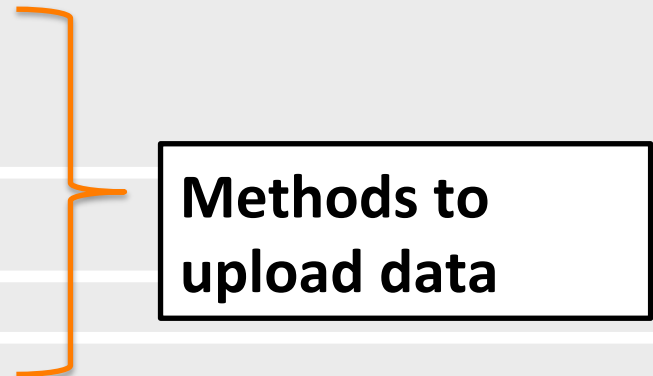
Name for this job (optional):

Input file format:




Either paste data:

Or upload file: No file selected.
Click [here](#) to download the previously uploaded file.

Or provide file URL:



Recent jobs

Show/hide columns (1 hidden)		Filter
Analysis	Jobs	Submitted at
Assembly Converter	 Assembly conversion of Oglab_chr12 in Oryza_glaberrima Queued	30/11/2016, 00:49 (GMT)
Assembly Converter	 Assembly conversion of Oglab_chr12 in Oryza_glaberrima Done [Download results]	29/11/2016, 16:59 (GMT)
Assembly Converter	 Assembly conversion of Oglab in Oryza_glaberrima Failed	29/11/2016, 16:51 (GMT)

Species supported for assembly mapping

Species:	Oryza sativa Japonica
Assembly mapping:	
Name for this job (optional):	
Input file format:	

- Arabidopsis thaliana
- Brachypodium distachyon
- Hordeum vulgare
- Oryza glaberrima
- Oryza sativa Indica
- Oryza sativa Japonica
- Populus trichocarpa
- Solanum lycopersicum
- Triticum aestivum
- Vitis vinifera
- Zea mays**

Chain files provided at ensemble FTP

ftp://ftp.ensemblgenomes.org/pub/release-33/plants/assembly_chain/

Convert GFF/GTF format files

- [GFF](#) (General Feature Format) is plain text file used to describe gene structure
- [GTF](#) (Gene Transfer Format) is a refined version of GFF

Sample GFF:

```
10   gramene gene      130112  134842  .       +       .       ID=gene:Zm00001d023208
10   gramene gene      218781  220384  .       -       .       ID=gene:Zm00001d023210
10   gramene gene      265243  271813  .       -       .       ID=gene:Zm00001d023211
10   gramene lincRNA_gene  271217  272132  .       -       .       ID=gene:Zm00001d023035
```

- Each feature (exon, intron, UTR, etc.) is processed separately and independently
- Plain text, compressed plain text, and URLs pointing to remote files are all supported
- Only chromosome and genome coordinates are updated

Convert BED format files

- [BED](#) (Browser Extensible Data) file is a tab-delimited text file describing genome regions or gene annotations

Sample BED:

chr	start	end
1	1630	1760
1	1982	9815
1	10217	11435
1	10371	11284
1	11720	14685

- It consists of one line per feature, each containing 3-12 columns (chr, start, end are required)
- Regions of old assembly which to multiple locations on new assembly will be split
- Compressed remote files are not supported!

Limitations

- Size limit 50 MB
- Need to convert chromosome IDs to numerical chromosome numbers (*i.e.* strip off the prefix)
 - *e.g.* OglabChr01 -> 1
 - OglabChr02 -> 2
- Does not support VCF, BAM, BigWig & Wiggle format yet

Demo

Converting GFF coordinates

IRGSP1 => MSU6

Input

1	agi	ncRNA_gene	2631	2760	.	-	.	ID=gene:EPLOSAG00000002326
1	<u>irgsp</u>	gene	2983	10815	.	+	.	ID=gene:OS01G0100100
1	<u>irgsp</u>	gene	11218	12435	.	+	.	ID=gene:OS01G0100200
1	<u>irgsp</u>	gene	11372	12284	.	-	.	ID=gene:OS01G0100300
1	<u>irgsp</u>	gene	12721	15685	.	+	.	ID=gene:OS01G0100400

Output

1	agi	ncRNA_gene	1631	1760	.	-	.	ID=gene:EPLOSAG00000002326
1	<u>irgsp</u>	gene	1983	9815	.	+	.	ID=gene:OS01G0100100
1	<u>irgsp</u>	gene	10218	11435	.	+	.	ID=gene:OS01G0100200
1	<u>irgsp</u>	gene	10372	11284	.	-	.	ID=gene:OS01G0100300
1	<u>irgsp</u>	gene	11721	14685	.	+	.	ID=gene:OS01G0100400

Converting BED coordinates

IRGSP1 => MSU6

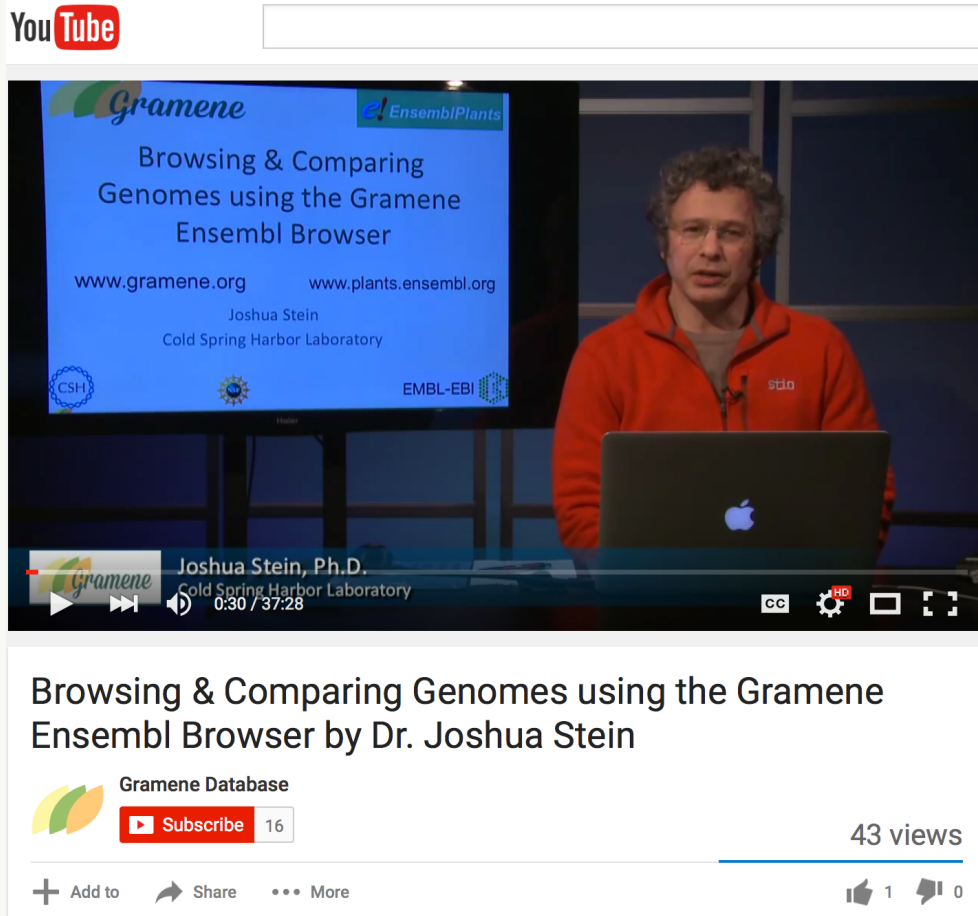
Input

1	2630	2760	gene:EPLOSAG00000002326	.	-	agi	ncRNA_gene	.	ID=gene:EPLOSAG00000002326
1	2982	10815	gene:OS01G0100100	.	+	irgsp	gene	.	ID=gene:OS01G0100100
1	11217	12435	gene:OS01G0100200	.	+	irgsp	gene	.	ID=gene:OS01G0100200
1	11371	12284	gene:OS01G0100300	.	-	irgsp	gene	.	ID=gene:OS01G0100300
1	12720	15685	gene:OS01G0100400	.	+	irgsp	gene	.	ID=gene:OS01G0100400

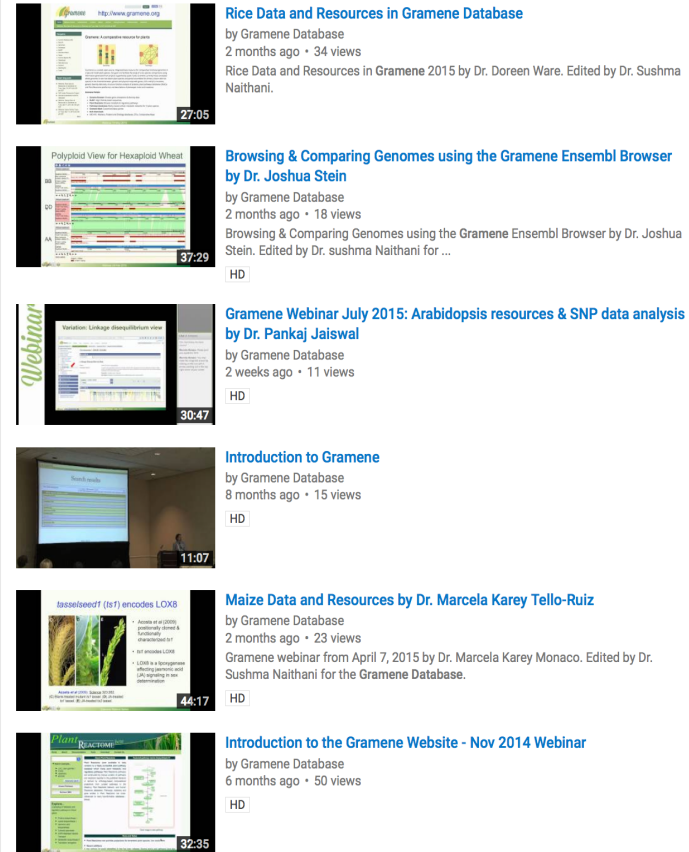
Output

1	1630	1760	gene:EPLOSAG00000002326	.	-	agi	ncRNA_gene	.	ID=gene:EPLOSAG00000002326
1	1982	9815	gene:OS01G0100100	.	+	irgsp	gene	.	ID=gene:OS01G0100100
1	10217	11435	gene:OS01G0100200	.	+	irgsp	gene	.	ID=gene:OS01G0100200
1	10371	11284	gene:OS01G0100300	.	-	irgsp	gene	.	ID=gene:OS01G0100300
1	11720	14685	gene:OS01G0100400	.	+	irgsp	gene	.	ID=gene:OS01G0100400

Gramene Database @ YouTube



The video player shows a man in an orange hoodie sitting at a desk with a laptop. Behind him is a large screen displaying a presentation slide. The slide has the Gramene and EnsemblPlants logos at the top. The main text on the slide reads: "Browsing & Comparing Genomes using the Gramene Ensembl Browser". Below this, it lists the website addresses "www.gramene.org" and "www.plants.ensembl.org", the name "Joshua Stein", and his affiliation "Cold Spring Harbor Laboratory". Logos for CSH, EMBL-EBI, and stio are also visible on the slide. The video player interface includes the YouTube logo, a search bar, and a video title: "Browsing & Comparing Genomes using the Gramene Ensembl Browser by Dr. Joshua Stein". Below the title is the channel name "Gramene Database" with a "Subscribe" button and "16" subscribers. The video has "43 views". At the bottom, there are icons for "Add to", "Share", and "More", along with a thumbs up icon showing "1" and a thumbs down icon showing "0".



A vertical list of video thumbnails and titles from the Gramene Database YouTube channel. Each entry includes a thumbnail, the video title, the uploader (Gramene Database), the upload date, and the view count. The videos are:

- Rice Data and Resources in Gramene Database** by Gramene Database, 2 months ago • 34 views. Rice Data and Resources in Gramene 2015 by Dr. Doreen Ware. Edited by Dr. Sushma Naithani. 27:05
- Browsing & Comparing Genomes using the Gramene Ensembl Browser by Dr. Joshua Stein** by Gramene Database, 2 months ago • 18 views. Browsing & Comparing Genomes using the Gramene Ensembl Browser by Dr. Joshua Stein. Edited by Dr. sushma Naithani for ... 37:29
- Gramene Webinar July 2015: Arabidopsis resources & SNP data analysis by Dr. Pankaj Jaiswal** by Gramene Database, 2 weeks ago • 11 views. 30:47
- Introduction to Gramene** by Gramene Database, 8 months ago • 15 views. 11:07
- Maize Data and Resources by Dr. Marcela Karey Tello-Ruiz** by Gramene Database, 2 months ago • 23 views. Gramene webinar from April 7, 2015 by Dr. Marcela Karey Monaco. Edited by Dr. Sushma Naithani for the Gramene Database. 44:17
- Introduction to the Gramene Website - Nov 2014 Webinar** by Gramene Database, 6 months ago • 50 views. 32:35

- 21 recorded webinars (topic specific or targeted to plant communities, *e.g.*, Arabidopsis, rice, maize, grape)
- Master mailing list of over 1,100 plant researchers

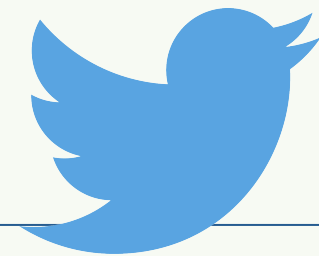
How to reach us

- E-mail: webinars@gramene.org
- Contact form:
<http://www.gramene.org/contact>
- Gramene YouTube channel
- Announcements mailing list



facebook

<https://www.facebook.com/Gramene>



[@GrameneDatabase](https://twitter.com/GrameneDatabase)

Gramene - Exploring Function through Comparative Genomics and Network Analysis

NSF IOS 1127112 (2011- 2017)

Doreen Ware, PI (USDA ARS, CSHL)

Michael Campbell, Kapeel Chougule, Yinping Jiao, Sunita Kumari, Andrew Olson, Joshua Stein, Marcela K. Tello-Ruiz, Peter van Buren, Bo Wang, Sharon Wei

Pankaj Jaiswal, Co-PI (OSU)

Noor Al-Bader, Justin Elser, Matthew Geniza, Parul Gupta, Justin Preece, Sushma Naithani

Paul Kersey / Robert Petryszuk (EMBL-EBI)

Dan Bolser, Christopher Grabmuller, Chuang Kee Ong, Dan Staines, Brandon Walts / Elisabet Barrera, Maria Keays, Oliver Mannion, Nuno Fonseca, Laura Huerta Martinez

Lincoln Stein (OICR)

Peter D' Eustachio (NYU); Guanming Wu, Robin Haw, Joel Weiser, Sheldon McKay; Antonio Fabregat (EBI)

Crispin Taylor (ASPB)

Patty Lockhart; Weijia Xu (TACC), Amit Gupta(TACC)

