Genome Literacy Workshop

Elisabeth Busch-Nentwich & Pavol Kramár

Queen Mary University of London, UK



Why genome literacy?

- Saves time and effort
 - Ensembl collates a lot of published information
 - Knowing how to find that can (at least partially) replace trawling through papers
- Better experimental design for gene editing experiments
 - Understand gene structure
 - Understand splice variants and expression behaviour
- Download data without coding
 - Sequence, phenotypes, gene ontology annotations
- Customise genome browser
 - Switch on relevant tracks, upload your own data

Learning Outcomes

- Understand Ensembl as a database
 - basics of default data
 - \circ investigating homology
- Find and switch on optional features
 - find your gene and its associated data
- Download gene and genome data
 - \circ key tools to use
- Upload and display your own data

Workshop structure

- Lecture sections
 - Introduce background information
 - Show examples
 - Slides for future reference
- Small exercises
 - In lecture sections
 - I will point out when to switch to browser
- Large exercises
 - Designed to be more challenging
 - Cover all tools and features we introduce
- Your questions
 - \circ $\,$ Email me so that we can work through them on Day 2 $\,$

Workshop structure

Day 1

- Part 1
 - Zebrafish genome basics
 - Ensembl basics
 - Nomenclature (Gene names, IDs, etc...)
 - Assembly vs annotation
- Part 2
 - Configuring Ensembl tracks
 - Ensembl "Gene" view
 - Comparative genomics
- Homework
 - Email me your questions so that we can work through them on Day 2
 - Finish exercises

Day 2

- Part 3
 - BioMart
 - Other tools (Variant Effect Predictor, etc...)
 - Custom tracks
- Your questions

Part 1

- Zebrafish Genome Project
- Ensembl
- Finding your gene
- Gene name and IDs
- Manual and automatic annotation
- Ensembl "Region" view

Ensembl

- Most examples from **Ensembl** (we are biased!)
- Probably most widely used genome browser amongst zebrafish researchers
- **Primary source of zebrafish annotation** (UCSC imports Ensembl annotation)
- Currently Ensembl version 110 (July 18th)
- New releases 3 or 4 times / year
- Zebrafish **annotation largely static** between releases
- But **naming and homology** updated (+ new functionality)



Zebrafish Genome

- GRCz11 (danRer11) latest assembly, released in 2017
- Sequencing strategy:
 - \circ 90% clone by clone sequencing
 - High quality
 - 10% whole genome shotgun sequencing
 - Lower quality
 - Fills gaps between clones
 - Identified by accessions beginning with CABZ

			60.83 kb			Forward strand
40.76M	40.7	7Mb 40.78Mb	40.79Mb	40.80Mb	40.81Mb	
htra3a-204 protein cor	4 - ENSDART00000189780 > ding			gpr78a-201 - ENSDART0000011 protein coding	11656 >	
htra3a-203 protein cor	1 - ENSDART0000080563 > ding				cpz-202 - ENSDART00000187394 > protein coding	2
D htra3a-203 protein cor	3 - ENSDART00000163584 > ding				cpz-201 - ENSDART00000147497 > protein coding	
C S	D^0- gpr78a-202 - ENSDART00000175582 > protein coding				cpz-203 - ENSDART00000193568 > protein coding	
CA820110	003791 > CAB201060264.1 >			CR387992.11 >		
40.76M	40.7	7Mb 40.78Mb	40.79Mb	40.80Mb	40.81Mb	

Zebrafish Genome History

- Genome project started in **2001** at Sanger Institute
- Initially sequenced pool of **Tübingen** zebrafish
- But zebrafish **very polymorphic** compared to humans
- Too much variation to join clones, so lots of **gaps**
- + same region represented by 2+ clones, leading to **artificial duplication**
- Later used **double haploid** Tübingen fish for some clones and most WGS
- Only 925 gaps between scaffolds and N50 > 7 Mbp
- GRCz11 contains alternative scaffolds
- When downloading sequence from Ensembl FTP site, "toplevel" includes alternative sequence, but "primary_assembly" doesn't and is probably what you want



Older Assemblies

- Previous assemblies available in Ensembl archives: www.ensembl.org/info/website/archives/assembly.html
 - GRCz10 / danRer10: <u>http://e91.ensembl.org/</u>
 - Zv9 / danRer7: <u>http://e77.ensembl.org/</u>
 - Zv8 / danRer6: <u>http://e54.ensembl.org/</u>
- Even older assemblies available in UCSC
- Numbering coordinated when GRC (Genome Reference Consortium) took over managing zebrafish assembly from Sanger Institute



Ensembl Mirrors

- Mirrors: <u>www.ensembl.org/info/about/mirrors.html</u>
- Main site (UK): <u>www.ensembl.org</u>
- US East mirror: useast.ensembl.org
- Most often slow due to chosen tracks though



• Follow link from **ZFIN**

= 🗳ZFIN	Search	Q. Sign In
dmd	GENE	
Summary	dmd	
Expression	ID	ZDB-GENE-010426-1
Phenotype	Name Symbol	dystrophin dmd Nomenclature History
Mutations	Previous	cb664 (1), Dp71 (1), Duchenne muscular dystrophy (1), im 6911785 san sanje like (1) sanje zf0YS (1) zar 110165
Human Disease	Type	protein coding gene 2
Gene Ontology	Location	Chr: Mapping Details/Browsers
Protein Domains	Description	Predicted to have actin binding activity and zinc ion binding
Transcripts	•	activity. Involved in several processes, including sarcomere organization; skeletal muscle organ development; and somatic
Interactions and Pathways		muscle development. Localizes to sarcolemma. Used to study Duchenne muscular dystrophy and muscular dystrophy. Human
Antibodies		ortholog(s) of this gene implicated in cognitive disorder; dilated
Plasmids		dystrophy (multiple). Is expressed in several structures,
Constructs		including axial mesoderm; axis; chordo neural hinge; musculature system; and somite. Orthologous to human DMD
Marker Relationships		(dystrophin).
Sequences	Genome Resources	Alliance 2 (1), Gene:837732 (1), VECA 01100000000 (5052 (1),
Orthology	Note	Ensembl(GRCz11):ENSDARG0000008487@(11) None

- Follow link from **ZFIN**
- Search by gene name on Ensembl (or old name or mutant name)

		Login/Registe
Ensembl BLA	ιST/BLAT VEP Tools BioMart Downloads Help & Docs Blog 🛛 🧧 🗸 Search Zebrafish	(
lew Search Jobs 🔻		
Current selection:		
< all Species	Only searching Zebrafish 🔻 dmd	
Only searching Zebrafish	21 results match dmd when restricted to species: Zebrafish 💥	
Restrict category to:	dmd (Zehrafish Gene)	
Gene 2	ENSDARG00000008487 1:10824351-11075405:-1 Dystrophin [Source:ZFIN:Acc:ZDB-GENE-010426-1]	
Transcript 13	dmd-201 (ZFIN transcript name record; description: dystrophin.) is an external reference matched to	
GeneTree 1	Transcript ENSDART0000007013	
GenomicAlianment 5	Variant table • Phenotypes • Location • External Refs. • Regulation • Orthologues • Gene tree	
	dmd (Zebrafish Alternative sequence Gene)	
Per page:	ENSDARG00000115779 CHR_ALI_C1G1_1_4:1099/816-11031921:-1 Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1]	
10 25 50 100	dmd-213 (ZFIN transcript name record; description: dystrophin,) is an external reference matched to Transcript ENSDART00000164141	
Layout:	Not a Primary Assembly Gene	
Standard Table	Variant table • Phenotypes • Location • External Refs. • Regulation • Gene tree	
otandara habio	dmd-212 (Zebrafish Transcript)	
Tip:	Dystrophin [Source:ZFIN;Acc:ZDB-GENE-010426-1].	
Help and Documentation can be	Location • External Refs. • cDNA seq. • Exons • Variant table • Protein seq. • Population • Protein summary	
searched from the homepage! Just	dmd-201 (Zebrafish Transcript)	
type in a term you want to know moi about like non-synonymous SNP	Fe ENSDART0000007013 1:10824653-10914523:-1 Dystrophin [Source: ZEIN: Acc: ZDB-GENE-010426-1]	
assac, and non synonymous one.	Location • External Refs. • CDNA seq. • Exons • Variant table • Protein seq. • Population • Protein summary	

- Follow link from **ZFIN**
- Search by gene name on Ensembl (or old name or mutant name)
- Search using **BLAST** or **BLAT** on Ensembl
 - BLAT is faster
 - BLAST finds more distant alignments + alternative scaffolds
 - No BLAST/BLAT on Ensembl archive sites but can use BLAT on UCSC

									Login/Regist
Ensemb BAS	T/BLAT VEP Tools BioMart	Downloads H	telp & Docs E	log	E	 Search Zeb 	rafish		
BLASTIBLAT Y									
Veb Tools 6 Web Soola 8 BLAST/BLAT	Results for dmd Ø								
B Totel	Job details 8								
	Job name	1							
- File Champleon	Species	Zebrafe	sh (Danio rerio)						
 Assembly Converter D Mature Converter 	Assembly	GRCz11							
VCF to PED Converter Data Sicar Post OWAS	Search type	BLASTN (N	(CBI Blast)						
O Configure this page	de Download results lie	Harw J.D							
A Custom tracks	Results table 😑								
🗄 Econara	Show Al v antries	Sho	while columns	(2 hidden)			Course		6
< Share this page	Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score "	E- val	1110
Bookmark Die page	CHR. ALT. CTG1, 1, 4:11031622- 11032521 (Issuence)	and	Reverse	1	900	900 (Internal	1779	0.0	100.000 (Alte
	5.11031622-11032521 (Sevenue)	and	Reverse	1	900	900	1779	0.0	100.000 1012
	1.1926324-1626215 (Instance)	chote	Reverse	1	387	390 (Sectore)	686	0.0	96.923 (blics
	1.11349264-11349645 (Semanar)	adk.th	Reverse	1	387	390 (Descent)	686	0.0	96.923
	5:45265567-45266368 (belowing)	EY15L	Reverse	1	387	390 (Instants)	685	0.0	96.923

- Follow link from **ZFIN**
- Search by gene name on Ensembl (or old name or mutant name)
- Search using **BLAST** or **BLAT** on Ensembl
 - BLAT is faster
 - BLAST finds more distant alignments + alternative scaffolds
 - No BLAST/BLAT on Ensembl archive sites but can use BLAT on UCSC
- Check gene correct by checking orthologues and/or synteny

Species	Type	Orthologue T	arget %/d	Query %id	GOC Score	WGA Coverage	High Confidence	٠
Japanese	1-to-many	dmd (ENSOR: 00000020638)	84.52 %	88.22 %	0	95.81	Yes	
(Oryzies latipes)	View Gene Tree	Compare Regions (2:208,119-221,155	1)					
		Yew Sequence Alignments						
Lumpfish	1-to-many	dmd (ENSCLMG0000509931)	82.21 %	82.49 %	0	95.20	Yes	
(Cyclopterus (umpus)	View Gene Tree	Contourn Regions (2:5,248,684- 5,281,983-1)						
		View Sequence Algoments						
Lyretail cichlid	1-10-1	dmd (ENSNERG0000015200)	87.34 %	89.39 %	0	96.66	Yes	
brichardi)	View Gene Thee	Correary Reviews (JH422367.1:2,004, 2,028,054)-1)	027-					
		View Sequence Aligoments						
Makobe Island	140-1	dmd (EMSPN()G000000226(1)	45.32 %	89.55 %	0	96.75	Yes	
(Pundamilia /yerersi)	View Gene Tree	Concern Regions (JH419417.1.620,20 712,305~1)	05-					
		View Sequence Alignments						

Gene Names

- Names assigned to Ensembl genes automatically based on sequence similarity
 - Mistakes are possible
 - Names can change
- **ZFIN gene symbols** (i.e. the name assigned by ZFIN) are preferred, but other databases are also used, e.g. HGNC (HUGO* Gene Nomenclature Committee) and miRBase
- Description indicates source of name
- Genes without a match are given a name based on the sequence used to identify them, e.g AL645792.1 (clone) or **CABZ**01052570.1 (WGS)

*HUGO: Human Genome Organization



- Best to use stable IDs
- e.g. **ENS**DARG0000028213 (ttn.2 or ttna)
- ENS = Ensembl

- Best to use stable IDs
- e.g. ENS**DAR**G0000028213 (ttn.2 or ttna)
- ENS = Ensembl
- **DAR** = Danio rerio

- Best to use stable IDs
- e.g. ENSDARG0000028213 (ttn.2 or ttna)
- ENS = Ensembl
- **DAR** = Danio rerio
- **G** = Gene (also T for Transcript, P for Peptide and E for exon)

- Not completely stable, if annotation or underlying assembly changes
- Stable IDs have versions, e.g. ENSDARG00000058767.4
 - Version number of **ENSDARG** increases if transcripts change
 - Version number of **ENSDART** increases if splicing, chromosome or sequence of transcript change
 - Version number of **ENSDARP** increases if peptide's sequence changes
 - Version number of **ENSDARE** increases if exon's sequence changes
- Can also be **removed**, e.g. searching for ENSDARG00000058767

- Not completely stable, if anno
- Stable IDs have **versions**, e.g.
 - Version number of ENSDARG ind
 - Version number of ENSDART inc change
 - Version number of ENSDARP inc
 - Version number of ENSDARE inc
- Can also be **removed**, e.g. sea



- Not completely stable, if annotation or underlying assembly changes
- Stable IDs have versions, e.g. ENSDARG00000058767.4
 - Version number of **ENSDARG** increases if transcripts change
 - Version number of **ENSDART** increases if splicing, chromosome or sequence of transcript change
 - Version number of **ENSDARP** increases if peptide's sequence changes
 - Version number of **ENSDARE** increases if exon's sequence changes
- Can also be **removed**, e.g. searching for ENSDARG00000058767
- Can use <u>www.ensembl.org/Danio_rerio/Tools/IDMapper</u> to convert older IDs to what they **map** to currently in Ensembl
- **Relevance**: Gene IDs from older publications!

- Not completely stable, if annotation or underlying assembly changes
- Stable IDs have versions, e.g. ENSDARG00000058767.4
 - Version number of **ENSDARG** increases if transcripts change
 - Version number of **ENSDART** increases if splicing, chromosome or sequence of transcript change
 - Version number of **ENSDARP** increases if peptide's sequence changes
 - Version number of **ENSDARE** increases if exon's sequence changes
- Can also be **removed**, e.g. searching for ENSDARG00000058767
- Can use <u>www.ensembl.org/Danio_rerio/Tools/IDMapper</u> to convert older IDs to what they **map** to currently in Ensembl

Mini-exercise: Check whether your favourite genes have had stable ID changes. What's the ENSDARG version number? (Hint: look for your gene's ID history)

Gene Annotation

- Zebrafish (+ human, mouse, rat) has **manual** and **automatic** gene annotation
- Other **300+** genomes in Ensembl only have automatic annotation
- www.ensembl.org/info/about/species.html



From Ensembl training materials, CC BY 4.0 license

Manual Annotation

- Gold standard
- Uses information from databases and publications
- More accurate for tricky areas:
 - e.g. UTRs, splice sites, single exon transcripts
- Slower and more expensive
- Thorough, but leads to inclusion of transcripts that may not be representative (e.g. low expression)
- Only clones manually annotated



From Ensembl training materials, CC BY 4.0 license

Automatic Annotation

- Faster
- Uses evidence from sequences deposited in ENA/GenBank/DDBJ and UniProt proteins

• Overview:

- Identify repeats and low complexity sequence with RepeatMasker, Dust and TRF
- Run GENSCAN to identify *ab initio* gene predictions
- Align UniProt proteins to GENSCAN predictions, prioritising zebrafish proteins or those from closely related or well annotated species
- Make gene models using Genewise
- Align cDNAs, ESTs and RNA-seq to annotate UTRs and make RNAseq gene models
- Collapse redundant transcripts and cluster into genes, prioritising manual annotation but including automatic annotation if different splicing
- Identify pseudogenes by looking for genes with frameshifts / repeats
- Identify processed pseudogenes by looking for multi-exon equivalent



From Ensembl training materials, CC BY 4.0 license

Merged Annotation

- Golden: Identical manual and automatic annotation
- Red: **Protein-coding** transcript from automatic annotation
- Blue: Non-coding transcript
- Filled box: Coding exon
- Non-filled box: Non-coding exon



 In reality, would not trust these retained intron transcripts unless shown to have comparable expression levels

Which Transcript?

- Often multiple transcripts
- **Best** transcript for experiments?
- Golden transcript is a good bet
- Ensembl Canonical transcript is, on balance, most conserved, most expressed, longest CDS (coding sequence) and in other databases
- APPRIS combines protein structure, important residues and homology to identify a principal isoform - APPRIS P1

Description		BRISC and BRCA1 A complex member 1 [Source:NCBI gene;Acc:445296] zgc:100909								
Gene Synonyms										
Location		Chrompsome 11: 6.051.287-6.070.192 reverse strand. GRCz11:CM002895.2								
About this gene		This gene has 4 transcripts (solice variants) and 185 orthologues.								
		_								
Transcripts		Hide	transcript (able						
Transcripts Show/hide columns (1 hi	dden)	Hide	transcript (able			Filter			
Show/hide columns (1 hi Transcript ID	dden) Name 💧	Hide	Protein	Biotype	UniProt Match		Filter			
Show/hide columns (1 hi Transcript ID ENSDART00000122262.3	dden) Name 🍦 babam1-202	Hide bp () 2035	Protein 370aa	Biotype	UniProt Match	0	Fitter Flags Ensembl Canonical	APPRIS P		
Show/hide columns (1 hi Transcript ID ENSDART00000122262.3 ENSDART00000008980.8	dden) Name babam1-202 babam1-201	Hide bp () 2035 1888	Protein (370aa 370aa	Biotype Protein coding Protein coding	UniProt Match Q6AXK4 19 A0A0R419A4 19 Q6AX	¢ (K4.52	Filter Flags Ensembl Canonical APPRIS I	APPRIS P		
Transcripts Show/hide columns (1 hi Transcript ID ENSDART00000122262.3 ENSDART0000006980.8 ENSDART000000162776.2	Name babam1-202 babam1-201 babam1-203	Hide bp () 2035 1888 802	Protein 370aa 370aa 197aa	Biotype Protein coding Protein coding Protein coding	UniProt Match Q6AXK4@ A0A0R4I9A4@Q6AX A0A0R4IJK1@	0 (K4-12)	Filtar Flags Ensembl Canonical APPRIS I CDS 3' incon	APPRIS P		

"Region in detail" Demo

• Go to "22:3153000-3217000"



- 4 clones
- 2 genes on +
- 1 gene on -
- Manual + automatic annotation
- cDNA + EST tracks
- Variant tracks

"Region in detail" Demo

• Go to "22:3153000-3217000"



- 4 clones
- 2 genes on +
- 1 gene on -
- Manual + automatic annotation
- cDNA + EST tracks
- Variant tracks

Quiz time!

Exercise 1

- Do Exercise 1 "exploring the genome"
- Covers:
 - $\circ \quad \text{Region view} \quad$
 - BLAST/BLAT
 - Archive sites
- Go to mbl2023.buschlab.org

Part 1 summary

- Zebrafish genome basics
- Ensembl basics
- Nomenclature (Gene names, IDs, etc...)
- Assembly vs annotation
- Learning outcomes:
 - Understanding gene features for experiments, e.g. CRISPR, ISHs
 - Distinguish between annotation and assembly, gene versions
 - Being able to switch between releases
 - Judging which transcripts/ gene features are reliable
- Any questions?

Part 2

- Configuring Ensembl tracks
- Ensembl "Gene" view
- Comparative genomics

 But first, back to the region we were looking at before the exercises: "22:3153000-3217000"

"Configure this page" Demo

• Go to "22:3153000-3217000" and click "Configure this page"

												LO	givikegister
	μ	ISEIIIDI BLAST/BLAT	/EP Tools	BioMart	Downloads H	elp & Docs Blog	_				 Search Zebri 	Mish	<u>م</u>
- Ze	۴.	Configure Region Image Configure	Overview Imag	ge Configu	re Chromosome In	hage Personal Data							
	-	Active tracks		Select from	n available coof	ourations:	Current une	nund		for a construction of a sector			
Location:		Favourite tracks		Select Iron	n avanable com	gurations.	Current uns	aveu		Save current comparation			
Location-b		Track order		Genes a	and transcr	ints	Coheman	tel.					
- Whole g		Genome Reference Consortium I	(07)	0011000	ina nanovi	ip to	Canow into						
- Region		Genome Reference Consortium in	(9/7)	RNASeq n	nodels								
- Region	ĥ	Sequence	(2/4)										
B Compar	1	Simple features	(0/3)	Filter by	All classes	~							_
- Aligni	Β	Genes and transcripts	(7/96)		Enter terms to	fiter by							_
- Aligni	I t	Genes Production transcripts	(2/2)										
- Regio	h	RNASeq models	(5/93)	Key	Shown	Hidden N	o Data	Elltored	Shown	Hidden			
- Varia	8	mRNA and protein alignments	(2/4)	nuy	Shown			Filtered.	Shown	nioden			
- Rese	H	mRNA alignments	(2/3)				-89						
- Markers		Protein alignments	(0/1)				-OTED						
Other g	Ĩ	Sequence variants	(2/8) (1/2)			ab ^{it}	annin's						
- UCS	H	Failed variants	(0/1)			Siles moo	Ra						
- NODE	l t	Phenotype annotations	(0/2)			of the cone offor							
Conf		Comparative secondar	(1/3)	Default st	yle:	ŇĒŇ							
a cust	Ĩ	Multiple alignments	(0/2)		~								
TE CON	H	Conservation regions	(1/2)	ENA									
🛃 Expo		BLASTZ/LASTZ alignments	(0/64)	1 dpf sam	ple1	0,1,0,							
- a		Oligo probes	(0/41)	14 dof sar	mple1	0 1 0							
< shar		Repeat regions	(1/23)	2 4 4 4 4 4 4									
P Book		Display options	(12/14)	z uprisam	pier	1 1 1							H+
		onspilay options		3 dpf sam	ple1	1 1 1							_
		Q Search for track hubs		5 dpf sam	ple1	1 1 1							0
				-		0 0 0			64.00 kb				
					3.1646	3.1716		3.18%6		3.1946	3.2046	3.2146	
		CDNA.	lignment							Manga da			

RefSeq Aside

- NCBI's **annotated** and **curated** database of reference sequences, including transcripts and proteins
- Accessions starting **X** are "Model RefSeq" **predictions** from automatic genome annotation
- Accessions starting N are "Known RefSeq" from manually curated cDNA and EST data
- Accessions starting NM & XM indicate mRNA; NP & XP are proteins
"Configure this page" Demo

• Go to "22:3153000-3217000" and click "Configure this page"



"Configure this page" Demo

- Go to "1:10822281-10882903" and click "Configure this page"
- Under "RNASeq models", turn on "Intron-spanning reads" for "pharyngula prim 5" and "pharyngula prim 15"



"Gene" Demo - Summary

• Go to ENSDARG00000102765

Gene-based displays		
🖻 Summary	Gene: lonp1 ENSDARC	G00000102765
 Splice variants Transcript comparison 	Description	Ion peptidase 1, mitochondrial [Source:ZFIN;Acc:ZDB-GENE-030131-4006
Gene alleles	Gene Synonyms	fc64d11, prss15, wu:fc64d11
Secondary Structure Comparative Genomics Concerning allocaments	Location	Chromosome 22: 3.160,447-3.182,965 reverse strand. GRCz11:CM002906.2
- Gene tree	About this gene	This gene has 2 transcripts (splice variants), 190 orthologues and 1 paralogue.
 Gene gain/loss tree Orthologues Paralogues 	Transcripts	Show transcript table
Ensembl protein families	Summary @	
 GO: Cellular component GO: Biological process 	Name	lonp1 @ (ZFIN)
GO: Molecular function	Ensembl version	ENSDARG00000102765.2
Phenotypes Genetic Variation	Gene type	Protein coding
- Variant table	Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.

"Gene" Demo - Transcript Table

• Go to ENSDARG00000102765 and click on "Show transcript table"

Gene-based displays	Gene: lonp1 ENSD/	ARG000001	102765	ŝ						
Splice variants Transcript comparison Gene alleles Sequence	Description Gene Synonyms	Description Gene Synonyms			lon peptidase 1, mitochondrial [Source:ZFIN;Acc: <u>ZDB-GENE-030131-4006</u> #] fc64d11, prss15, wu:fc64d11					
Secondary Structure Comparative Genomics Genomic alignments	Location			Chromosome 22: 3.160.447-3.182.965 reverse strand. GRCz11:CM002906.2						
 Gene tree Gene gain/loss tree Orthologues Paralogues 	About this gene Transcripts		Hide transcript table							
Contologies	Show/hide columns (1 hi	dden)					Filter			
- GO: Cellular component - GO: Biological process - GO: Molecular function	Transcript ID ENSDART00000158009.2	Name lonp1-201	bp 4114	Protein d	Biotype Protein coding	UniProt Match	Flags Ensembl Canonical	APPRIS P1		
Phenotypes Genetic Variation	ENSDART00000167550.2	lonp1-202	741	<u>247aa</u>	Protein coding	A0A0R4IPW4@	CDS 5' and 3' in	complete		
- Variant table - Variant image - Structural variants	Summary @									
 Gene expression 	Name		loop	1@ (ZFIN)						
- Pathway - Regulation	Ensembl version		ENS	DARGOOOD	00102765.2					
- External references	Gene type		Prof	ein coding						
Supporting evidence ID History	Annotation method		Ann	otation for t	his gene includes l	both automatic anno	tation from Ensembl a	and Havana manual curation, see		

"Gene" Demo - Splice Variants

• Go to ENSDARG00000102765 and click on "Splice variants"



"Gene" Demo - Orthologues

• Go to ENSDARG00000102765 and click on "Orthologues"

Show All 🛩 entries		Show/hide columns				Filter	
Species	Туре	a Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Abingdon island giant	1-10-1	LONP1 (ENSCAB/G0000010924)	75.03 %	75.26 %	25	n/a	No
(Chelonoidis abingdonii)	View Gene Tree	Company Regions (PKMU01001122.1:170,198- 221,187:1)					
		View Sequence Alignments					
African ostrich (Struthio complue	1-10-1	LONP1 (ENSSCUG0000004632)	80.69 %	70.50 %	50	n/a	Yes
australis)	View Gene Tree	Compare Regions (KL206174.1:174,870-201,335:1)					
		View Sequence Alignments					
Algerian mouse (Mercanetus)	1-to-1	Lonp1 (MGP_SPRETELJ_G0022694)	75.05 %	74.12 %	0	n/a	No
(wins shumps)	View Gene Tree	Compare Regions (17:54,555,161-54,567,779:-1)					
		View Sequence Alignments					
Alpine marmot	1-to-1	LONP1 (ENSMMMG00000018859)	62.80 %	62.22 %	0	n/a	No
(marmota)	View Gene Tree	Conteare Regions (CZRN01000089.1:3,499,006- 3,520,539:-1)					
		View Sequence Alignments					
Amazon molly	1-to-1	lonp1 (ENSPEC/G0000001826)	75.94 %	77.43 %	0	85.71	Yes
(Policina formosa)	View Gene Tree	Compare Regions (KI520250.1:178,559-209,184:-1)					
		View Sequence Alignments	0005				

"Gene" Demo - Paralogues

• Go to ENSDARG00000102765 and click on "Paralogues"

Gene: lonp1 ENSD	ARG000001	102765	5							
Description		lon	peptidase 1.	, mitochondrial [So	urce:2FIN;Acc.2	DB-GENE-030131-400	147			
Gene Synonyms		1064	d11. prss15	i, wu:fc64d11						
Location		GR	hromosome 22: 3.160.447-3.162.965 reverse strand. RCz11:CM002906.2							
About this gene		This	gene has 2	transcripts (splice	variants), 190 c	thologues and 1 paralo	GLE.			
Transcripts		F	fide transcrip	pt table						
Show/hide columns (1 h	idden)					Filter				
Transcript ID 0	Name	bp ()	Protein ()	Biotype	UniProt Match	6 Flags	0			
ENSDART00000158009.2	lonp1-201	4114	966aa	Protein coding	A0A0R4IH798	Ensembl Canonica	APPRIS P1			
ENSDART00000167550.2	lonp1-202	741	24788	Protein coding	A0A0R4IPW4	CDS 5' and 3'	ncomplete			
Download paralogue Show/hide columns	3								Filler.	
Туре	Ancestral ta	xonor	ny Ense	mbl identifier & g	jene name (Compare	Location		Target %id	Query %id
Paralogues	3ilateral anir (Bilateria)	nals	ENS lonp2 lon pe	DARG0000010143 2 ptdase 2. peroxisomal Acc 4940301	(Source:NC8I	Region Companison Alignment (protein) Alignment (cDNA)	18:18:475:574-18	524.624 - 1	36.31 %	31.57 %

"Gene" Demo - GO Terms

• Go to ENSDARG00000102765 and click on "GO: Molecular function"

GO: Molecular	function Ø				
Show/hide columns	(1 hidden)			Filte	ar 🛅
Accession	Term	Evidence	Annotation source	Transcript IDs	¢
<u>GO:0000166</u> @	nucleotide binding	IEA	UniProt	ENSDART00000158009	Search BioMart View on karyotype
<u>GO:0003677</u>	DNA binding	IEA	UniProt	ENSDART00000158009	Search BioMart View on karyotype
<u>GO:0003697</u>	single-stranded DNA binding	IBA	GO_Central	ENSDART00000158009	Search BioMart View on karyotype
<u>GO:0004176</u> @	ATP-dependent peptidase activity	IBA	GO_Central	ENSDART00000167550 ENSDART00000158009	Search BioMart View on karyotype
<u>GO:0005524</u>	ATP binding	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	Search BioMart View on karyotype
<u>GO:0016887</u>	ATP hydrolysis activity	IEA	UniProt	ENSDART00000158009 ENSDART00000167550	Search BioMart View on karyotype
GO:0043565@	sequence-specific DNA binding	IEA	UniProt	ENSDART00000158009	Search BioMart View on karyotype

"Gene" Demo - External References

• Go to ENSDARG00000102765 and click on "External references"

External references @									
This gene corresponds to the following database identifiers:									
		Filter							
External database	Database identifier								
Expression Atlas	ENSDARG00000102765 dP [view all locations]								
NCBI gene (formerly Entrezgene)	lonp1成 Ion peptidase 1, mitochondrial [view all locations]								
WikiGene	lonp1d Ion peptidase 1, mitochondrial [view all locations]								
ZFIN	lonp1d과 Ion peptidase 1, mitochondrial <u>[view all locations]</u>								

"Gene" Demo - Expression Atlas

• From "External references" click "Expression Atlas" ID then "18 White et al"

White RJ, C	• White RJ, Collins JE, Sealy IM, Wali N, Dooley CM et al. (2017) A high-resolution mRNA expression time course of embryonic development in zebrafish.								
Raw Data Provider: Vertebrate Genetics and Genomics Group (Wellcome Trust Sanger Institute)									
Results	Experiment Design	Supplementary Information	Downloads						
Genes		Show boxplot and transcripts view	1						
ENSDARG0000	0102765 ×	Showing 1 gene:						Ensembl genome browser -	Ownload
LICONICOUD						Click on a cell t	o open the select	ed genome browser with attach	ed tracks if available
		Expression level in TPM 0	179						
Apply	Clear			and the	and h particular	Stranger and and	r	the second second	dback
Most specific	c	a spalate in the	a wat a bare with	Street is stilled	Sheet allowing in	stan in station To	a phine sha phine	displan strate and a	Lee de
Expression va	alue	right distant parties	the the state	Charling Charling	and the and the	and the party	Party Party	and and and	used -
0.5	-	lonp1							

Compara

- Compara produce Ensembl's comparative genomics resources
- Two types of analysis:
 - Gene level comparisons to produce gene trees, e.g. infer homologues (orthologues & paralogues)
 - Whole genome alignments pairwise and multiple alignments, e.g. constrained elements and synteny

Compara - Gene Trees

- Separate trees for **proteins** and **ncRNAs** (take secondary structure into account)
- Process:
 - Take **representative** transcripts (e.g. longest CDS) from all genes from all species
 - Classify genes into **clusters** by TreeFam family
 - Build **multiple** alignment
 - Build **gene tree** reconciled with NCBI's taxonomy tree
 - Infer orthologues and paralogues



Compara - Infer Homologues (Orthologues & Paralogues)



z1 & z2 are paralogues (arose from duplication), as are c1 & c2

z1 & c1 are orthologues (arose from speciation), as are z2 & c2 + z2 & g, etc...

z1 & c1 have a one-to-one relationship

g has a one-to-many relationship to e.g. z1 and z2

Homologues labelled "high confidence" are supported by conservation of synteny or whole genome alignment blocks

Compara - lonp1 Gene Tree



Mini-exercise: Go to lonp1 gene tree and expand/collapse subtrees to show both paralogs

- Due to an apparent production issue, many fewer genes in release 110 have **ZFIN** names
- 23,328 genes in release 109 have ZFIN names, but only 15,924 now
- Meanwhile, only **164** genes in 109 had HGNC names, but now **3,113** do
- (308 genes have names from miRBase in both 109 and 110)
- Expect fix, but not until next release at the earliest (i.e. 3 or 4 months)
- Affects all types of protein-coding genes, but a particular problem for **paralogs...**

Teleost-specific paralogs are normally called *a* and *b*, e.g. *kdm2aa* and *kdm2ab* Paralogs for *kdm2aa* in Ensembl release 109:

Show/hide colum	ns				Filter		
Туре	Ancestral taxonomy	Ensembl identifier & gene	Compare	Location		Target %id	Query %id
Paralogues	Osteoglossocephalai	ENSDARG0000078133 kdm2ab lysine (K)-specific demethylase 2Ab [Source:ZFIN;Acc:ZDB-GENE-101007- 5]	Region Comparison Alignment (protein) Alignment (cDNA)	<u>14:21,754,521-21,782,227:1</u>		52.40 %	53.62 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000046010 kdm2bb lysine (K)-specific demethylase 2Bb [Source:ZFIN;Acc:ZDB-GENE-080225- 13]	Region Comparison Alignment (protein) Alignment (cDNA)	<u>10:41,826,331-41,907,213:-1</u>		42.70 %	43.32 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000036593 kdm2ba lysine (K)-specific demethylase 2Ba [Source:ZFIN;Acc:ZDB-GENE-040426- 2195]	Region Comparison Alignment (protein) Alignment (cDNA)	<u>8:40,238,647-40,276,193:-1</u>		29.45 %	20.93 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000018559 kdm7ab lysine (K)-specific demethylase 7Ab [Source:NCBI gene;Acc:503902]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>25:17,244,532-17,310,301:1</u>		22.20 %	16.43 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000006584 phf8 PHD finger protein 8 [Source:NCBI gene,Acc.566534]	 Region Comparison Alignment (protein) Alignment (cDNA) 	23:27,591,366-27,608,257:-1		20.74 %	17.23 %

Teleost-specific paralogs are normally called a and b, e.g. *kdm2aa* and *kdm2ab* Paralogs for *kdm2aa* in Ensembl release 110:

Show/hide col	umns				Filter	
Туре 🔶	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id 🔻	Query %id
Paralogues	<u>Osteoglossocephalai</u>	ENSDARG0000078133 KDM2A lysine demethylase 2A [Source:HGNC Symbol;Acc:HGNC:13606]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>14:21,754,521-21,782,227:1</u>	55.07 %	56.36 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000046010 KDM2B lysine demethylase 2B [Source:HGNC Symbol;Acc:HGNC:13610]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>10:41,826,331-41,907,213:-1</u>	42.22 %	42.83 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000036593 kdm2ba lysine (K)-specific demethylase 2Ba [Source:ZFIN;Acc:ZDB-GENE-040426-2195]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>8:40,238,647-40,276,193:-1</u>	29.33 %	20.85 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000018559 kdm7ab lysine (K)-specific demethylase 7Ab [Source:ZFIN;Acc:ZDB-GENE-050309-32]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>25:17,244,532-17,310,301:1</u>	21.65 %	16.02 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG0000006584 phf8 PHD finger protein 8 [Source:NCBI gene (formerly Entrezgene);Acc:566534]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>23:27,591,366-27,608,257:-1</u>	20.64 %	17.15 %

Teleost-specific paralogs are normally called a and b, e.g. *kdm2aa* and *kdm2ab* Paralogs for *kdm2aa* in Ensembl release 110:

Show/hide colu	Filter					
Туре 🔶	Ancestral taxonomy	Ensembl identifier & gene name	Compare	Location	Target %id 🔻	Query %id
Paralogues	Osteoglossocephalai	ENSDARG00000078133 KDM2A lysine demethylase 2A [Source:HGNC Symbol;Acc:HGNC:13606]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>14:21,754,521-21,782,227:1</u>	55.07 %	56.36 %
Ancient paralogues	Animals and Fungi (Opisthokonta)	ENSDARG00000046010 KDM2B lysine demethylase 2B [Source:HGNC Symbol;Acc:HGNC:13610]	 Region Comparison Alignment (protein) Alignment (cDNA) 	<u>10:41,826,331-41,907,213:-1</u>	42.22 %	42.83 %

Three problems:

- Makes it unclear that a gene is a paralog
- Violates nomenclature rules
- Breaks link to ZFIN genes

Species	Gene	Protein
Zebrafish	shha	Shha
Human	SHH	SHH
Mouse	Shh	SHH

Accessing previous release:

 Variant table Variant image 	Namo	clc24c5-f1 (ZEINI					
L Structural variants	Encombl vorgion	SIC2483EP (A						
 Gene expression Pathway 	Case type	Distain and	0000024771.0					
 Molecular interactions 	Gene type	Protein cod	ng in shin nana inaludan kash automatin an	- station from Encounted and Decement				
 Regulation External references 	Annotation method	Annotation	or this gene includes both automatic an	notation from Ensembliand Havana m	anual curation, see <u>anticle</u> .			
- Supporting evidence	Go to Reg	jion in Detail for more tracl	s and navigation options (e.g. zoomi	ing)				
└─ Gene history								
🌣 Configure this page	Add/remove tracks	🛓 Custom tracks < Sha	re 🕀 Resize image 🗖 Export imag	e 🎭 Reset configuration 👼 Rese	t track order			
Custom tracks					34.08 kb			Forward strand
	Gapor	5.205Mb	5.210Mb	5.215Mb	5.220Mb	5.225Mb	5.230Mb	5.235Mb
Export data	(Comprehensive set)			slc24a5-201 - ENSDART0000003 protein coding	3574 >			
 Share this page 	Contigs	C4	BZ01080601.1 >			CU457753.8 >		
🛃 Bookmark this page	Genes (Comprehensive set)	< golim4a-203 - ENSDART000 protein coding	00183109					< myef2-201 - ENSDART0000008: protein coding
								< myef2-203 - ENSDART0000009 protein coding
								<pre></pre>
		5.205Mb Reverse strand	5.210Mb	5.215Mb	5.220Mb 34.08 kb	5.225Mb	5.230Mb	5.235Mb
	Gene Legend	Protoin Codina						
		Ensembl protein coding						
		merged Ensembl/Hava	na					
	Configuring the di	splay						
	Tip: use the "Configure	e this page" link on the left to	show additional data in this region.					
	Ensembl release 110 - July 2	2023 © <u>EMBL-EBI</u>						Permanent link: View in archive site
	About Us		Get help		Our sister sites		Follow us	
	About us		Using this website		Ensembl Bacteria		Blog	



Accessing previous release:

CENSembl BLA	ST/BLAT VEP Tools BioMart Downloads Help & Docs Blog	Login/Register ☑ - Search all species Q
Using this website Annotation and	prediction Data access API & software About us	
In this section Discrete Section Control Panel Find a Data Display	Help & Documentation	
Adding Custom Tracks Track Hubs Tutorials Glossary Supported browsers Archives Ametation a Prediction Ensembl Stable IDs Variation Comparative Genomics Regulation Microarray Probe Mapping Known Bugs Ensembl constantian	Using this website Our website offers lots of ways to view and interact with our genomic data - find out more! • Adding custom tracks • Tutorials • Glossary. • EAQs (Frequently Asked Questions)	Annotation & Prediction The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. • Ensembl annotation • Variation data • Comparative genomics • Regulatory annotation
Accessing Ensembl Data Exporting data via website API data access Public MySQL Server ETD Duraled	More	More
FIP Download BioMart Virtual Machine ApJ & Software	Data access All of our data is open-access and can be downloaded free of	API & Software Ensembl releases all its software under an Apache-style
 Ensembl Tools API Documentation Doxygen Perl documentation 	charge (<u>disclaimer</u>). Ways to access this data include: • Export features or sequence directly from web pages	open source licence. Our products include: Perl API for direct data access

Extract data from our public database using Perl

– DAS (Distributed Annotation Syst

ver for language agnectiv



Accessing previous release:



Compara - Whole Genome Alignments

- Pairwise whole genome alignments with LASTZ
- Zebrafish has alignments to 64 species (plus itself)
- Only human (180) and medaka (65) have more
- Full list at: <u>www.ensembl.org/info/genome/compara/analyses.html</u>
- Multiple genome alignments with EPO (Enredo, Pecan, Ortheus)
- Zebrafish is in 2 alignments (out of 11 in Ensembl) one of 39 fish and one of 65 fish
- For lists of species, see:

www.ensembl.org/info/genome/compara/multiple_genome_alignments.html

• No zebrafish orthologue listed for human RBM20 gene (ENSG00000203867)

O Species without orthologues

22 species are not shown in the table above because they don't have any orthologue with ENSG00000203867.

- Ancestral sequence
- · Siamese fighting fish (Betta splendens)
- Sloth (Choloepus hoffmanni)
- Channel bull blenny (Cottoperca gobio)
- Lumpfish (Cyclopterus lumpus)
- Tongue sole (Cynoglossus semilaevis)
- Common carp (Cyprinus carpio carpio)

Zebrafish (Danio rerio)

• If we look at the region around RBM20 in human and then click on **Synteny** we see conservation of synteny with zebrafish chr22

Homo sapiens genes	Location		Danio rerio homologues	Location	
DUSP5 (ENSG00000138166)	<u>10:110497907-110511533</u>	\rightarrow	dusp5 (ENSDARG00000019307)	22:29911326-29922872	Region Comparison
SMC3 (ENSG00000108055)	<u>10:110567684-110606048</u>	→	smc3 (ENSDARG00000019000)	22:29858535-29906764	Region Comparison
RBM20 (ENSG00000203867)	10:110644336-110839468		No homologues		
PDCD4 (ENSG00000150593)	10:110871795-110900006	\rightarrow	pdcd4b (ENSDARG00000041022)	22:29655981-29689981	Region Comparison
BBIP1 (ENSG00000214413)	10:110898730-110919201	\rightarrow	bbip1 (ENSDARG00000071046)	22:29648854-29652356	Region Comparison
SHOC2 (ENSG00000108061)	<u>10:110919367-111017307</u>	\rightarrow	shoc2 (ENSDARG00000040853)	22:29596646-29640181	Region Comparison
ADRA2A (ENSG00000150594)	<u>10:111077029-111080907</u>	\rightarrow	adra2a (ENSDARG00000040841)	22:29584800-29586608	Region Comparison

• If we look at the chr22 region in zebrafish then all the surrounding genes are the same and RBM20 is likely to be BX649294.1



Homo sapiens genes	Location		Danio rerio homologues	Location	
DUSP5 (ENSG00000138166)	10:110497907-110511533	094 8	dusp5 (ENSDARG00000019307)	22:29911326-29922872	Region Comparison
SMC3 (ENSG00000108055)	10:110567684-110606048	+	smc3 (ENSDARG00000019000)	22:29858535-29906764	Region Comparison
RBM20 (ENSG00000203867)	10:110644336-110839468		No homologues		
PDCD4 (ENSG00000150593)	10:110871795-110900006		pdcd4b (ENSDARG0000041022)	22:29655981-29689981	Region Comparison
BBIP1 (ENSG00000214413)	10:110898730-110919201	-	bbip1 (ENSDARG00000071046)	22:29648854-29652356	Region Comparison
SHOC2 (ENSG00000108061)	10:110919367-111017307		shoc2 (ENSDARG00000040853)	22:29596646-29640181	Region Comparison
ADRA2A (ENSG00000150594)	10:111077029-111080907		adra2a (ENSDARG0000040841)	22:29584800-29586608	Region Comparison

• Erroneously labelled as processed transcript and so not in protein gene tree, so not labelled as orthologue or named by orthology

Name	BX649294.1 (C	lone-based (Ensen	ibl) gene)			
Ensembl version	ENSDARG000	00092881.3				
Gene type	Processed tran	script				
Annotation method	Annotation for t	his gene includes b	oth automatic annotati	on from Ensembl and H	Havana manual curatio	n, see <u>article</u> .
Go to R	egion in Detail for more tracks a	nd navigation opt	ions (e.g. zooming)			
Add/remove tracks	Share	🕀 Resize image	Export image Q	Reset configuration	B Reset track order	Forward strand
Contigs Genes (Merged Ensembl/Havana)	29.700Mb	29.725Mb	29.750Mb	29.775Mb	29.800Mb	29.825Mb
	<pre> 14 < pdcd4b-202 - ENSDART000001 protein coding 14 </pre>	82173	8x649294.1-20 processed transcr	1 - ENSDART000001355 ipt	03	
	protein coding	09223				
	< 8X649294 1-204 - EN processed transcript	SDART00000180697	0			
	++0.40.40.40.41.41.42.41.42.41.42.41.42.41.42.41.42.41.42.41.42.41.42.41.42.41.42.41.41.41.41.41.41.41.41.41.41.41.41.41.	SEART00000125017	0			
	< 8X649294.1-202 - EN processed transcript	ISDART00000190613	0 1			
	29 20045	29 72556	29 750Mb	29.775Mb	29 800Mb	29 825Mb

Exercise 2

- Do Exercise 2 "exploring genes"
- Covers:
 - \circ Gene view
 - Phenotypes
 - Gene Ontology
 - \circ Homologues
 - \circ Gene trees
 - o Synteny
- Go to mbl2023.buschlab.org

Summary overview of Ensembl Browser



Part 2 summary

- Switching on additional tracks
- Gene view
- Comparative genomics
- Learning outcomes:
 - Customise Ensembl tracks for your needs
 - Navigate gene view to find splice variants, exon tables, external data, etc
 - Identify orthologs and paralogs
- Any questions?
- Email me any questions that you think of later: e.busch-nentwich@qmul.ac.uk

Part 3 (Day 2)

- BioMart
- Other tools
- Custom tracks

• Before we start: Any questions from yesterday?

BioMart

- Export (large amounts of) Ensembl data without programming
- Completely **customisable**, but **simple** to make complex queries
- Four stages:
 - Dataset
 - $\circ \quad \text{Filters}$
 - Attributes
 - \circ Results

CENSEMBI BLAST	Login/Regis	ster Q	
> New Count Results	THE DESCRIPTION OF Help		
Dataset [None selected]	- CHOOSE DATABASE - ~		

BioMart - Dataset

• Choose database (e.g. genes or variants) and species

🤉 New 📓 Count 🗐 Results	TURL O XML	Part 🕐 Help
Dataset	Ensembl Genes 107 V	
[None selected]	CHOOSE DATASET - Chicken (maternal Broiler) genes (bGalGal1.mat.broiler.GRCg7b) Human genes (GRCh38.p13) Mouse genes (GRCm39) Rat genes (mRatBN7.2) Zebrafish genes (GRC211)	
	Abingdon island giant tortoise genes (ASM359739v1) African ostrich genes (ASM69896v1) Algerian mouse genes (SPRET_EiJ_v1) Alpaca genes (vicPac1) Alpine marmot genes (marMar2.1) Amazon molly genes (Poecilia_formosa-5.1.2) American bison genes (Bison_UMD1.0) American black bear genes (ASM334442v1) American black bear genes (ASM334442v1) American mink genes (NNQGG.v01) Arabian camel genes (CamDro2) Arctic ground squirrel genes (ASM342692v1) Argentine black and white tegu genes (HLtupMer3) Armadillo genes (Dasnov3.0)	

BioMart - Filters

- Filter to reduce the dataset
- Can select **multiple** filters
- e.g. regions, IDs, GO terms, etc...

New Count Results		🛉 URL 👂 XML 🗿 Peri 🛞 Help
Dataset Zebrafish genes (GRCz11)	Transcript count >= Transcript count <=	1
Filters Chromosome/scaffold: 22 Start: 300000 End: 4000000 Transcript count <=: 1 Gene type: protein_coding	Gene type	polymorphic_pseudogene processed_pseudogene processed_transcript protein_coding pseudogene
Attributes Gene stable ID Gene stable ID version Transcript stable ID Transcript stable ID version	Transcript type	antisense IG_C_gene IG_C_pseudogene IG_J_pseudogene IG_pseudogene IG_V_pseudogene
Dataset	Source (gene)	ensembl ~
[None Selected]	Source (transcript)	ensembl ~
	APPRIS annotation	Only Excluded

BioMart - Attributes

- What data to **export**
- e.g. IDs, genomic locations, sequences, homologues, etc...

🥥 New 📄 Count 📓 Results		╈ URL 💿 XML 📑 Peri 💿 Help
Dataset 6 / 37241 Genes Zebrafish genes (GRCz11)	Please select colu Missing non coding g	mns to be included in the output and hit 'Results' when ready genes in your mart query output, please check the following \underline{FAQ}
Filters Chromosome/scaffold: 22 Start: 3000000 End: 4000000	Features Varia Structures Seq Homologues (Max select 6 orthologues)	ant (Germline) uences
Transcript count <=: 1	B GENE:	
Gene type: protein_coding Attributes Gene stable ID Gene name Source of gene name APPRIS annotation Chromosome/scaffold name Gene start (bp) Gene end (bp) Strand	Ensembl Gene stable ID Gene stable ID version	APPRIS annotation
	Transcript stable ID Transcript stable ID version Protein stable ID Protein stable ID Exon stable ID Gene description	Readthrough Gene name Source of gene name Source of transcript name Transcript count
	Chromosome/scaffold name	Gene % GC content Gene type
Dataset [None Selected]	Strand	Source (gene)

BioMart - Results

- Access your selected data in multiple formats
- e.g. HTML, TSV, CSV, XLS

🤊 New 📓 Count 🔳 Results	🚖 URL 🔁 XML 🗊 Peri 🕐 Help							
Dataset 6 / 37241 Genes Zebrafish genes (GRCz11) Filters	Export all results to Email notification to		File		/ ✓ □ Unique results or ML /	ily 🥝 Go		
Chromosome/scaffold: 22 Start: 3000000	View		10 v rows as HTML v	Unique re XLS	y y			
End: 4000000	Gene stable ID	Gene name	Source of gene name	APPRIS annotation	Chromosome/scaffold name	Gene start (bp)	Gene end (bp)	Strand
Transcript count <=: 1	ENSDARG00000103139	LO017843.1	Clone-based (Ensembl) gene	principal1	22	3045495	3078347	1
Gono tupo: protoin, coding	ENSDARG00000100132	CU929402.1	Clone-based (Ensembl) gene	principal1	22	3232925	3234494	1
Gene type: protein_cooling	ENSDARG00000100533	si:ch1073-178p5.3	ZFIN	principal1	22	3238474	3239834	1
Attributes	ENSDARG00000110077	CU929402.2	Clone-based (Ensembl) gene	principal1	22	3244950	3271707	1
Gene stable ID	ENSDARG00000053074	gipc3	ZFIN	principal1	22	3303671	3328241	1
Gene name	ENSDARG00000104717	tbxa2r	ZFIN	principal1	22	3336723	3344613	-1
Source of gene name APPRIS annotation Chromosome/scaffold name Gene start (bp) Gene end (bp) Strand								
Exercise 3

- Do Exercise 3 "CRISPR design"
- Covers:
 - Using BioMart to download sequence
 - Designing and checking guideRNAs
- Go to mbl2023.buschlab.org

More Tools

						Log Plage at
Ensembl BLAST	TIBLAT VEP Tools BioMart C	Downloads Help & Docs Blog		C1 - Search	all species	C
Jsing this website Annotation and p	rediction Data access API & soft	ware About us				
n this section	+ Help & Docomentation API &	L Software Ensembli Tools				
Ensembl Variant Effect Predictor VEP web interface VEP command line	Ensembl Tools					
- Data formats - Variant Recoder - Haplosaurus	We provide a number of ready-made to save the results indefinitely.	cels for processing both our data and yours. We routinely delete results from our t	servors after 10 da	iys, but if you have	ensembl.account	you will be able to
VEP FAQ Variant Simulator	Processing your data					
TOT IN FED CONTINUE	Name	Description	Online tool	Upload limit	Download script	Documentation
Search 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*	¢	0
	Variant Recoder	Translate a variant identifier, HGVS notation or genomic SPDI notation to all possible variant IDs, HGVS, VCF format and genomic SPDI.	`	Maximun 1000 variants recommended	¢	0
	BLAST/BLAT	Search our genomes for your DNA or protein sequence.	*	SOMB		0
	File Chameleon	Convert Ensemblifies for use with other analysis tools	*		cb	0
	Assembly Converter	Map (liftover) your data's coordinates to the current assembly.	*	50MB		0
	ID History Converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	*	50MB	ch	0
	Linkage Disequilibrium Calculator	Calculate LD between variants using genotypes from a selected population.	*			0
	VCF to PED converter	Parse a vcf file to create a linkage pedigree file (ped) and a marker information file, which together may be loaded into id visualization tools like Haploview.	*		eb.	0
	Data Slicer	Get a subset of data from a BAM or VCF file.	*			0
	Post-GWAS	Upload GWAS summary statistics and highlight likely causal gene candidates.	*		ф	0

• Results from all tools can be stored indefinitely if you create an **Ensembl account**

Variant Effect Predictor

- VEP predicts consequences of variants
- <u>www.ensembl.org/Danio_rerio/Tools/VEP</u>
- Example:
 - 22 3169475 3169475 G/T 1
 - 22 3169514 3169514 A/T 1
 - 22 3166910 3166910 C/A 1

(Chr, Start, End, REF/ALT, Strand)

• Custom Ensembl format, but standard formats like VCF can be used

Variant Effect Predictor



Exercise 4

- Do Exercise 4 "exploring variation"
- Covers:
 - Variant nomenclature
 - Variant Effect Predictor
 - Variant tables
- Go to mbl2023.buschlab.org

Assembly Converter

- Assembly Converter allows converting coordinates from one assembly to another
- Also known as LiftOver
- e.g. used for converting coordinates found in old papers
- www.ensembl.org/Danio_rerio/Tools/AssemblyConverter
- Example:
 - 22 3144711 3144711 sa39354
 - 22 3145013 3145013 sa43743
 - (Chr, Start, End, Name)
- BED format: www.ensembl.org/info/website/upload/bed.html
- (Only first three fields are essential)

Assembly Converter

Assembly Converter Ø		
New job		Clear form
This online tool currently uses CrossMap metadata in files, so track definitions, etc, v	which supports a limited number of formats (see our online documentation for details of the individu III be lost on conversion.	ual data formats listed below). CrossMap also discards
Species:	Zebrafish (Danio rerio)	
Assembly mapping:	GRCz10-> GRCz11 ~	
Name for this job (optional):		
Input file format:	BED ~	
Either paste data:	22 3144711 3144711 sa39354 22 3145013 3145013 sa43743	
Or upload file:	Choose file No file chosen	
Or provide file URL:		
	Runi	10

Assembly Converter

Assembly Convert	ter O					
New job						Clear form
This online tool currently use metadata in files, so track der	Inpu	ıt:			an ya walata barata kutoka na kata na anili wasan	CrossMap also discards
Species:	22	3144711	3144711	sa39354		
Assembly mapping: Name for this job (optiona	22	3145013	3145013	sa43743		
Input file format:						
Either paste data:	Out	put:				
	22	3161984	3161984	sa39354		
	22	3162286	3162286	sa43743		
Or upload file:			не словен			
Or provide file URL:						
				Run i	0	

UCSC In-Silico PCR

- Fast search for possible products from a pair of **PCR** primers
- genome.ucsc.edu/cgi-bin/hgPcr

UCSC In-Silico PCR

	–	ñ	Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data	Projects	Help	About Us
•	Fast searc	UCSC In	n-Silico PC	R							
•	genome.uc	Zebrafis	Genome: h	A	ssembly: RCz11/dani	Rer11) 🗸 🖸	Forward Prim	er: F TTTGA TGG	Reverse Prim	er: STCTG	submit
		Max P	roduct Size:	4000 Min P	erfect Mat	ch: 15	Min Good Ma	atch: 15	Flip Reve	rse Prime	r: 🗆
		A									
		About II	n-Silico PC	ĸ							
		In-Silico See an e	PCR searche example <u>vide</u>	es a sequence datal o on our YouTube cl	base with a hannel.	a pair of PC	R primers, using	an indexing	strategy for f	ast perfor	mance.
		Config	uration Op	tions							
		Genome Target - Forward	e and Assem If available, o Primer - Mu	bly - The sequence choose to query tran ist be at least 15 ba	database scribed se ses in leng	to search. equences. gth.	aar Minimum In	anth of 15 ha			
		Max Pro Min Peri Min Goo	duct Size - I fect Match - od Match - N	Maximum size of an Number of bases th umber of bases on	at match e at match e a of p	ion. exactly on 3' rimers when	end of primers.	Minimum ma of 3 bases ma	atch size is 1 atch.	5.	
		Flip Rev	erse Primer	- Invert the sequen	ce order o	f the reverse	e primer and cor	nplement it.			
		Output									
		When su between is capital example	iccessful, the and include lized in areas from human	search returns a se the primer pair. The where the primer s	equence or fasta hea equence r	utput file in f der describe natches the	fasta format con es the region in t database seque	taining all see the database ence and in lo	quence in the and the prim wer-case els	database ers. The f ewhere. H	e that lie asta body Here is an
		>chr22:3 TtACAGAT	1000551+310 TGATGATGCAT	01000 TAACAGATTGA GAAATGGGggggtggccag	TGATGCATG gggtggggg	AAATGGG CCC gtga	ATGAGTGGCTCCTA	AAGCAGCTGC			

UCSC In-Silico PCR

- Fast search for p
- genome.ucsc.ed

' n		Genomes	Genome Browser	Tools	Mirrors	Downloads	My Data	mers
M	UCSCI	n-Silico PCI	२					11010
J								
501	>chr14	KZ115440v1 al	t:182433-183230 79	8bp CCCG	GGAGCAGTTT	GAT CGTTGGGTGG	AGTAGGTCTG	
	CCCGGGG	AGCAGTTTGATCa	accttgctggaggtaagca	ictaaatcco	tct			
	tocataa	attgcatgctgct	ttcataactagatttgca	agagtttg	atg			
	gtgatat	tttagcgtcctcf	tcatttaaatataaagtt	atacatog	tgga			
	gttattt	gaataatgtgtat	aaataatattgcatcgat	gtaaagtaa	aaaa			
	tatcatt	taaattaaagcto	pacagcagttaatatggag	tcactgtaa	aag			
	cttaagga	atgaatgaagcat	ttaagagaatagcttcat	tttaaaaaco	cag			
	tgaaaaaa	gogatcaggato	cactaaqqataaaqaqac	totcoaga	aaa			
	tgcacat	gtagatggttaag	cttgcgctggctcttgtc	tctgaggaa	agat			
	cggcaaa	tgtgtgtgcgtgt	gtgtgtgtgtatgtgtggtt	tggtgggg	igta			
	capacto	cgctccatcagca	acctgctgctctgtctgag	acctctoo	aaa			
	atttata	tocactaaaataa	ctgagttaacaagtcatg	caatagtti	tcta			
	acccact	ctctttctgtgtf	tgtgcttgcttttgcaga	tacgcttte	aga			
	aatttcc	atggagaaaccto	tgtcagatCAGACCTACT	CCACCCAA	G	******		
	>chr14:	21334984-213: AGCAGTTIGATCI	5781 7980p CCCGGGG	AGCAGITIC	SAT CGTTGGG	TGGAGTAGGTCTG		
	tgattaa	attocatoctoct	ttcataactagatttgca	agagtttg	ato			
	tgcataa	aaatgctgacatt	taaataagtaataaatgt	gttatgata	aat			
	gtgatat	tttagcgtcctct	tcatttaaatataaagtt	atacatgg	tgga			
	gttattt	taataatgtgtat	taaataatattgcatcgat	gtaaagtaa	1000			
	ctaaqqa	atgaatgaagcet	ttaaqaqaataqcttcat	tttaaaaaco	cag			
	cttttca	gtcttttaaagtg	cattttgaataaatttaa	gctgtgcaa	atg			
	tgaaaaa	gggatcaggatgt	gagtaaggataaagagac	tgtcgagag	jaaa			
	tgcacat	gtagatggttaag	ottgcgctggctcttgtc	tectgaggaa	igat			
	cccttta	coctccatcanca	cctoctoctctotctoa	attotooo	igra			
	cagagtc	tcgacagccagca	ggaccccccaaataaatc	acctctgga	atc			
	atttata	tgcactaaaataa	octgagttaacaagtcatg	caatagtti	tcta			
	acccact	ctctttctgtgtt	tgtgcttgcttttgcaga	tacgcttte	aga			
	aatttee	atggagaaaccto	tgtcagattAGACCIACI	CLALCCAA	.0			

UCSC & Ensembl Differences

- Ensembl: 1 UCSC: chr1
- Ensembl: 1-based coordinates (bases numbered) UCSC: 0-based coordinates (numbers between bases)



• The **G** is **1:4-4** in Ensembl coordinates but **1:3-4** in UCSC

Custom Tracks

• Click "Custom tracks" and add <u>https://mbl2023.buschlab.org/data/3p-seq.bed</u>

orfigure Region Image Conf	Gorfgure Dverview Image Corfigure Dvomosome I	mage Personal Defa	
Frack Hub Registry Search	Your data		
Manage Configurations	Add a custom track		
	Please note that track hubs and indexed fill for more information.	es (BAM, BigBed, etc) do not work with certain cloud services, including Google Drive	and Dropbox. Please see our <u>succort.psop</u>
	Name for this data (optional):	3P-Seq	
	Species:	Zebrafish (Danio renio) Assembly: GRCz11	
	Data:	https://mbl2022.buschlab.org/data/3g- seq.bed	
		Or upload file (max 20MB) Choose file No file chosen	
	Data format:	BED v	
		Help on autoported formats, display types, etc	
		Add data	

• 24 hpf 3P-Seq data from Bartel lab

Custom Tracks

• Go to "22:3153000-3217000" (reverse strand)



Custom Tracks

• Go to "22:3153000-3217000" (forward strand)



- Lawson et al. (2020) "An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes" eLife 9:e55792
- Aim: To unify discrepancies between Ensembl and RefSeq annotations

 Lawson et al. (2020) "An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes" eLife 9:e55792



- Lawson et al. (2020) "An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes" eLife 9:e55792
- www.umassmed.edu/lawson-lab/reagents/zebrafish-transcriptome/
- Add:

https://www.umassmed.edu/globalassets/lawson-lab/downloadfiles/v4.3.2.gtf

• Large, so Ensembl will be slow - disable or delete when done



Regulatory information

- Ensembl integrates **regulatory information** from a variety of genome-wide assays (e.g. ChIP-seq and DNase-seq)
- But only has annotation for a limited set of species (mainly human and mouse, but also pig, chicken and some fish)
- Not zebrafish
- Instead have DANIO-CODE, a compendium of functional genomics datasets: <u>https://danio-code.zfin.org/</u>
- Datasets submitted to DANIO-CODE can be visualised in Ensembl (and UCSC)

Regulatory information

- For Ensembl, add a **custom track** using the DANIO-CODE track hub: <u>https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt</u>
- (You will need to select "**Track Hub**" for the data format.)
- Does make Ensembl quite slow, even with just the default tracks displayed
- Instead, could use UCSC by following instructions at:

https://danio-code.zfin.org/help/

Regulatory information



Exercise 5

- Do Exercise 5 "exploring data"
- Covers:
 - BioMart
 - $\circ \quad \text{Making BED files} \\$
 - Finding candidate genes
 - \circ Finding orthologues
- Go to mbl2023.buschlab.org

Part 3 Summary

- BioMart
- Other tools
- Custom tracks
- Learning outcomes:
 - Use BioMart to download data
 - Analyse variants
 - Making BED files
 - Upload custom tracks

Thank You!

Any questions?

