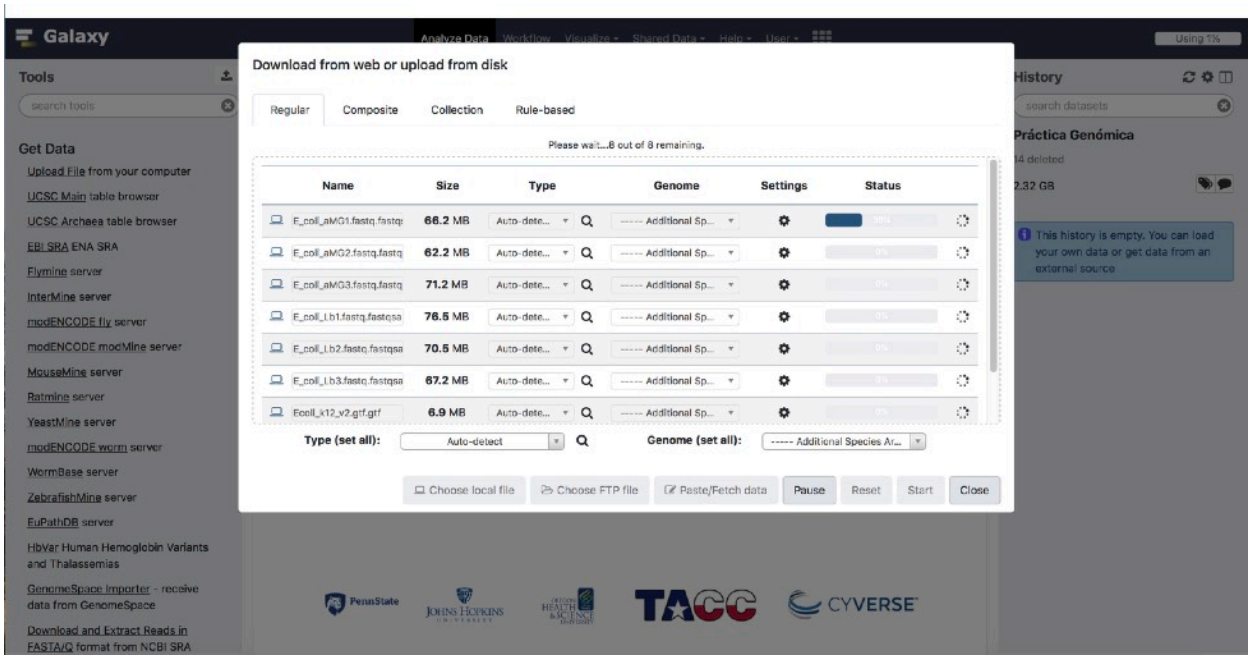
Aplicaciones de análisis

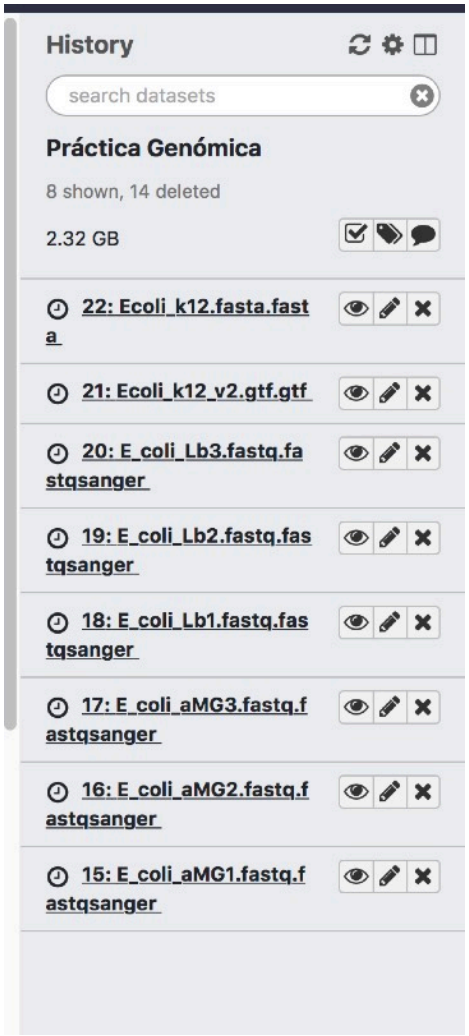
Práctica 1 - Genómica

José Manuel Bustos Muñoz

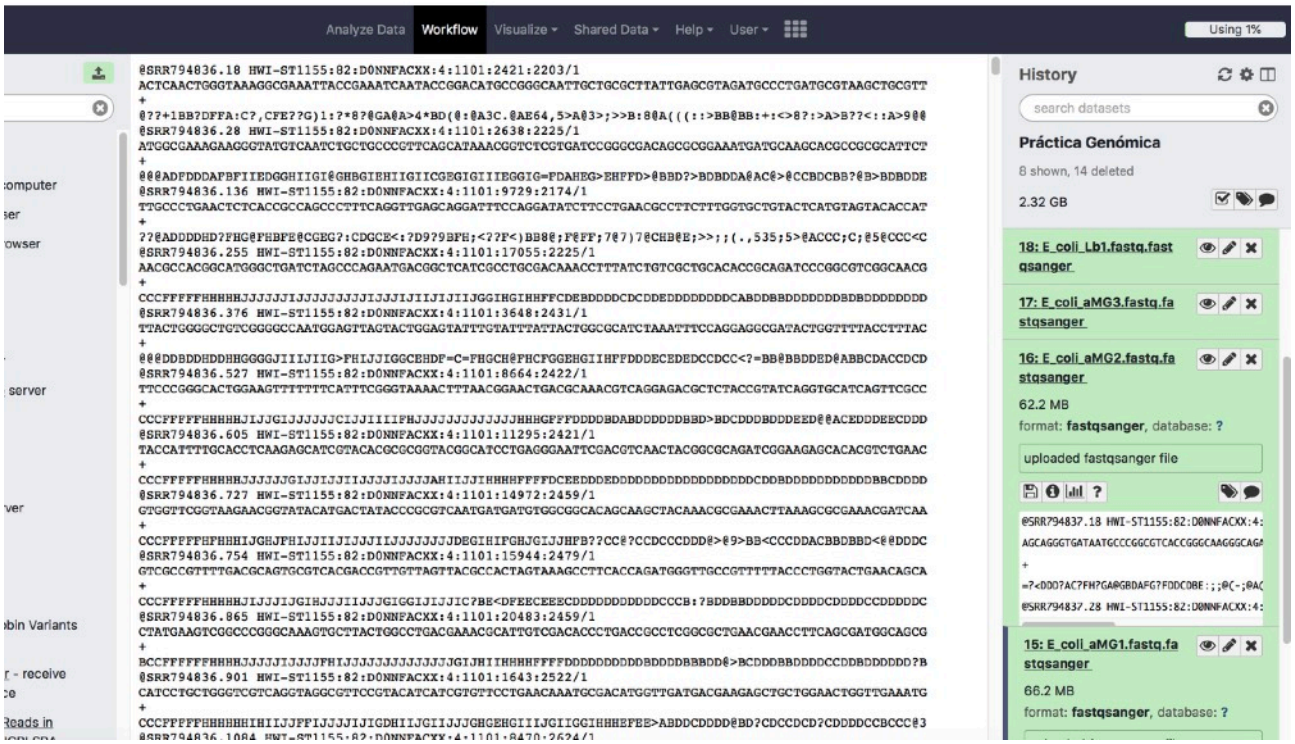
Lo primero sería realizar la carga de los datos para la práctica en la herramienta online de Galaxy.



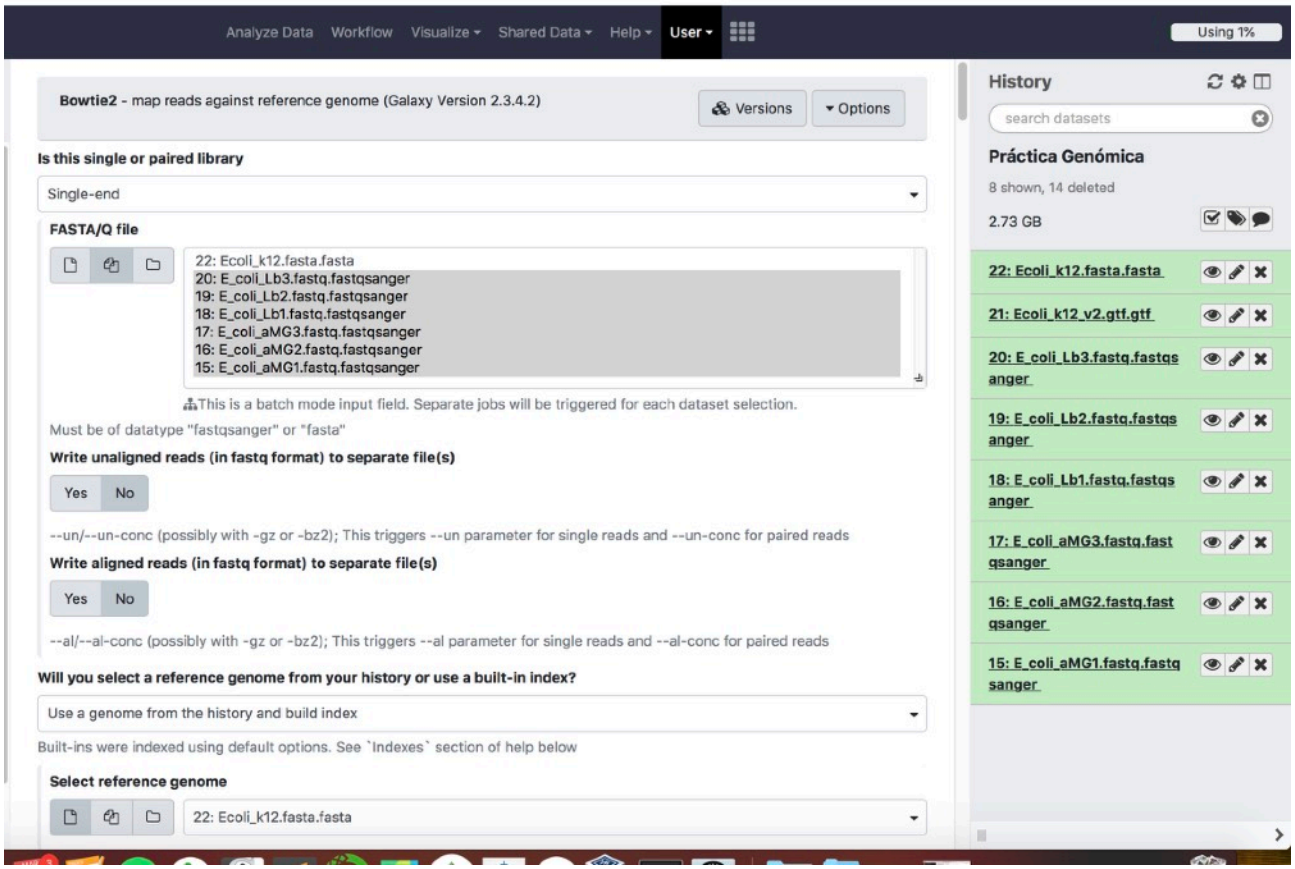
Una vez están cargados los datos nos aparece que archivo en el panel de History. Habría 3 ficheros fastq para cada una de las dos condiciones, un fichero fasta y un fichero gtf. Pueden visualizarse los datos de cada fichero o por ejemplo modificar el nombre o algunas propiedades similares.



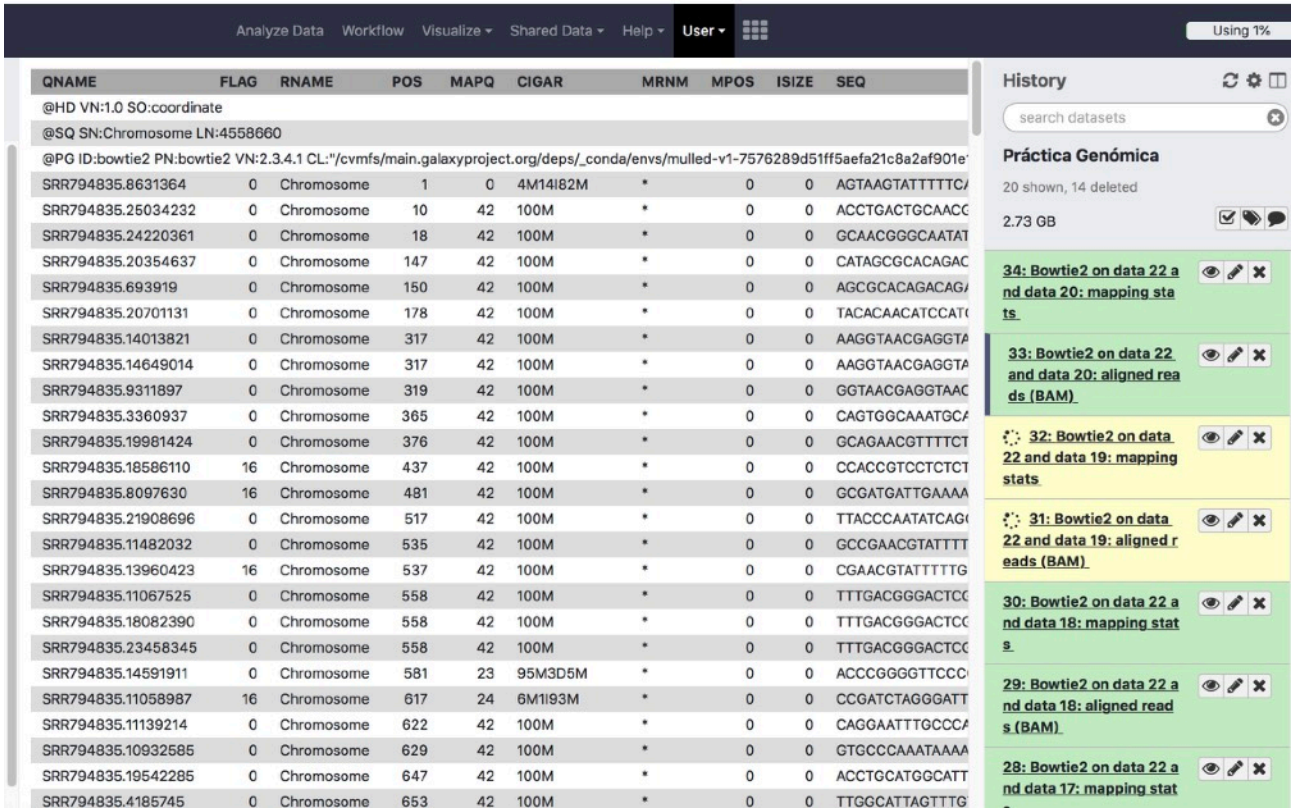
Visualización de datos de uno de los ficheros:



Ahora se alinean las lecturas de los ficheros “fastq” con el genoma de referencia que sería el fichero “fasta”. Para ello se ejecuta desde las tools la alineación seleccionando el tipo “NGS:Mapping —> Bowtie2”.



Se generan dos ficheros por cada uno de los 6 ficheros fastq que teníamos cargados. Por un lado se genera un fichero “aligned reads” que ve las alineaciones entre datos fastq y el genoma fasta:



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ
@HD VN:1.0 SO:coordinate									
@SQ SN:Chromosome LN:4558660									
@PG ID:bowtie2 PN:bowtie2 VN:2.3.4.1 CL:"/cvmfs/main.galaxyproject.org/deps/_conda/envs/mulled-v1-7576289d51ff5aefa21c8a2af901e									
SRR794835.8631364	0	Chromosome	1	0	4M14I82M	*	0	0	AGTAAGTATTTTTC/
SRR794835.25034232	0	Chromosome	10	42	100M	*	0	0	ACCTGACTGCAACC
SRR794835.24220361	0	Chromosome	18	42	100M	*	0	0	GCAACGGGCAATAT
SRR794835.20354637	0	Chromosome	147	42	100M	*	0	0	CATAGCGCACAGAC
SRR794835.693919	0	Chromosome	150	42	100M	*	0	0	AGCGCACAGACAG/
SRR794835.20701131	0	Chromosome	178	42	100M	*	0	0	TACACAACATCCAT
SRR794835.14013821	0	Chromosome	317	42	100M	*	0	0	AAGGTAACGAGGTA
SRR794835.14649014	0	Chromosome	317	42	100M	*	0	0	AAGGTAACGAGGTA
SRR794835.9311897	0	Chromosome	319	42	100M	*	0	0	GGTAACGAGGTAAC
SRR794835.3360937	0	Chromosome	365	42	100M	*	0	0	CAGTGGCAAATGC/
SRR794835.19981424	0	Chromosome	376	42	100M	*	0	0	GCAGAACGTTTTCT
SRR794835.18586110	16	Chromosome	437	42	100M	*	0	0	CCACCGTCCTCTCT
SRR794835.8097630	16	Chromosome	481	42	100M	*	0	0	GCGATGATTGAAAA
SRR794835.21908696	0	Chromosome	517	42	100M	*	0	0	TTACCCAATATCAG
SRR794835.11482032	0	Chromosome	535	42	100M	*	0	0	GCCGAACGTATTTT
SRR794835.13960423	16	Chromosome	537	42	100M	*	0	0	CGAACGTATTTTTG
SRR794835.11067525	0	Chromosome	558	42	100M	*	0	0	TTTGACGGGACTC
SRR794835.18082390	0	Chromosome	558	42	100M	*	0	0	TTTGACGGGACTC
SRR794835.23458345	0	Chromosome	558	42	100M	*	0	0	TTTGACGGGACTC
SRR794835.14591911	0	Chromosome	581	23	95M3D5M	*	0	0	ACCCGGGGTTCCC
SRR794835.11058987	16	Chromosome	617	24	6M1I93M	*	0	0	CCGATCTAGGGATT
SRR794835.11139214	0	Chromosome	622	42	100M	*	0	0	CAGGAATTTGCCCA
SRR794835.10932585	0	Chromosome	629	42	100M	*	0	0	GTGCCCAAATAAAA
SRR794835.19542285	0	Chromosome	647	42	100M	*	0	0	ACCTGCATGGCATT
SRR794835.4185745	0	Chromosome	653	42	100M	*	0	0	TTGGCATTAGTTTG

Y luego los ficheros “mapping stats” que contienen información estadística sobre el resultado de la alineación anterior.

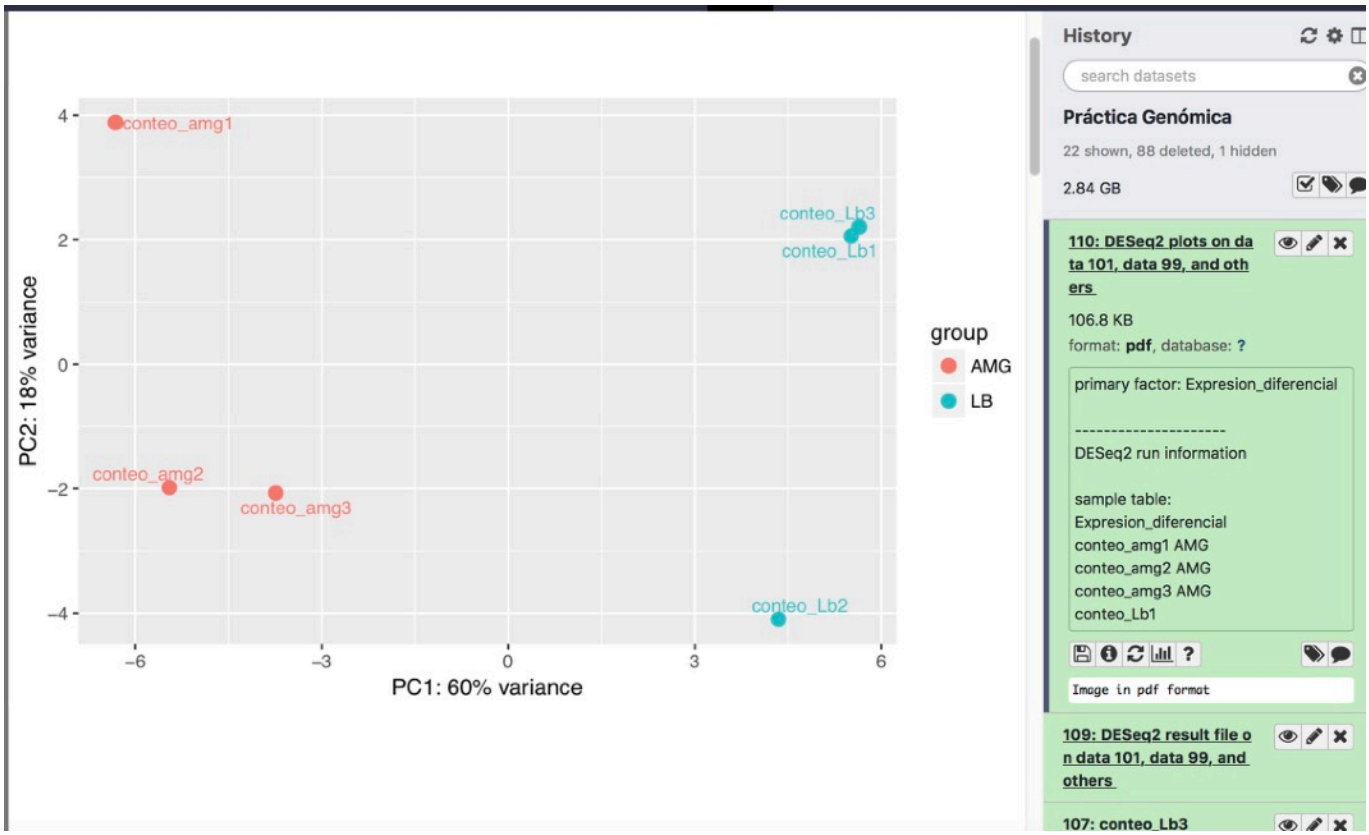
Borramos los ficheros de estadísticas, y nos centramos en los de alineación que renombramos para mayor comodidad.

Ahora se van a contar cuántas lecturas de los ficheros fastq encajan con cada gen. Para ello se usa el fichero GFF que contiene la información sobre la posición. Se hace el conteo desde la opción de Tools: “NGS:RNA Analysis - featureCounts”.
Nos quedamos con los 6 ficheros generados que contienen el conteo del número de veces que un gen se ha expresado. Renombramos los ficheros:

Geneid	alineacion_Lb3
Geneid	alineacion_Lb3
ER3413_4519	3
ER3413_1	93
ER3413_2	38
ER3413_3	32
ER3413_4	2
ER3413_5	18
ER3413_6	1
ER3413_7	311
ER3413_8	7
ER3413_9	7
ER3413_10	0
ER3413_11	1
ER3413_12	778
ER3413_13	49
ER3413_14	0

Con los ficheros que ya tenemos, se va a realizar el cálculo de la expresión diferencial: “NGS:RNA Analysis - DESeq2”, con el que se podrá determinar qué genes se encuentran diferencialmente expresados entre las condiciones del experimento. Se generan dos ficheros, uno que contiene diferentes gráficas que representan estadísticas, y otro fichero que contiene una fila para cada gen identificado en el fichero “gtf” y las métricas usadas para el cálculo de la expresión diferencial.

Damos un vistazo al fichero de gráficas:



Vemos el fichero generado para cada gen, con las métricas obtenidas:

GeneID	Base mean	log2(FC)	StdErr	Wald-Stats	P-value	P-adj
ptsG	287.104041489715	3.70386159302384	0.22148949706929	16.7225157040523	8.9845203743366e-63	2.02511089237547e-59
setA	119.192022574126	-3.77876481042538	0.284314606117701	-13.2907867872994	2.61810684587796e-40	2.95060641530631e-37
deoC	118.764760969725	-2.26082669697138	0.256851547594629	-8.80207543284686	1.343088543115 74e-18	1.00910719206096e-15
sucB	97.7830118745143	-2.04568679574465	0.244509440461329	-8.36649411934748	5.93578816712734e-17	3.34481663217626e-14
sucD	78.7277170967774	-2.18476604460764	0.265910784143029	-8.21616186665259	2.10119802553539e-16	9.47220069911353e-14
sodA	641.110578417242	-1.53279491780711	0.196485128179998	-7.80107345530464	6.13827679195282e-15	2.30594598151028e-12
deoA	74.6739675455924	-2.23164899075234	0.288814158356041	-7.72693763856697	1.10164579309581e-14	3.5472994537685e-12
ackA	492.668877883933	1.34270180231898	0.179911167478185	7.46313762030245	8.44860215855345e-14	2.380393655817244e-11
sucC	55.7557118136257	-2.19273413482056	0.304469130697969	-7.20182742268125	5.94108596092494e-13	1.48791197288054e-10
ibpA	33.1767454212724	2.34243447457831	0.339927483556994	6.89098289454892	5.54082018474337e-12	1.24890086964116e-09
sucA	81.1336674851828	-1.8763483981581	0.276691859978205	-6.78136464985238	1.190459765565736-11	2.43936028325922e-09
ydbK	174.490668050152	-1.73304373871055	0.265244787546982	-6.53375229250672	6.41419995866732e-11	1.20480055890301e-08
ompF	202.90284595257	1.36902711797722	0.210078135412546	6.5167520422283	7.18460426435596e-11	1.24569984706603e-08
yldA	54.5206411219009	-1.81678123811936	0.289059853928816	-6.28513857398806	3.27561377191371e-10	5.27373817278108e-08
nupG	88.034910150372	-1.68385950643439	0.275692357483031	-6.10774822271973	1.01046587133545e-09	1.51839338266006e-07
cydA	217.064331559983	1.21898657234768	0.204507133901558	5.96060660130543	2.51303300409762e-09	3.54023524452252e-07
ydjN	21.6418555368854	2.28898171915388	0.387857431070873	5.90160594018789	3.599801517952 29e-09	4.7729133067438e-07
gatZ	100.0378332595	-1.6631435918259	0.284637810078987	-5.84301710080041	5.126375275307 14e-09	6.41936103919016e-07
osmY	21.5449203180075	-2.19790739239054	0.384828700439431	-5.71139156170208	1.120560889132 58e-08	1.32933907584465e-06
gatC	46.8783496675536	-1.8055075323	0.319213792966511	-5.65610751190008	1.548444933567 91e-08	1.74509744013103e-06
icd	345.977927680529	-1.03786035495016	0.183887440333525	-5.64399805156756	1.661459542871 13e-08	1.78329990934835e-06
leuD	10.5406848202 43	-2.47238776541326	0.441851173212041	-5.59552155862848	2.199595119057 81e-08	2.25358518107105e-06
ahpC	418.033938553654	-0.970478463015697	0.177675922742626	-5.46207076364249	4.706124006799 99e-08	4.61200152666399e-06
aceE	1494.83214001797	0.862041733327824	0.161594197682971	5.33460820801901	9.575107442398 35e-08	8.99262173965245e-06
deoB	314.648375015497	-1.11975017428218	0.210326648487294	-5.32386258391705	1.015866604274 72e-07	9.15905330414087e-06
plaP	242.916594747527	1.05638020467497	0.205582468654296	5.13847416849228	2.769782590904 96e-07	2.31225554070362e-05
fadL	153.648740339467	1.0922137909467	0.212479086953045	5.14033548717229	2.742483533372 86e-07	2.31225554070362e-05

Preguntas:

1. ¿Qué genes han resultado diferencialmente expresados?

Nos quedamos con los genes con menor valor p-ajustado (tercera columna de la tabla):

ptsG	3.70386159302384	2.02511089237547e-59
setA	-3.77876481042538	2.95060641530631e-37
deoC	-2.26082669697138	1.00910719206096e-15
sucB	-2.04568679574465	3.34481663217626e-14
sucD	-2.18476604460764	9.47220069911353e-14
sodA	-1.53279491780711	2.30594598151028e-12
deoA	-2.23164899075234	3.5472994537685e-12
ackA	1.34270180231898	2.38039365817244e-11
sucC	-2.19273413482056	1.48791197288054e-10
ibpA	2.34243447457831	1.24890086964116e-09
sucA	-1.8763483981581	2.43936028325922e-09
ydbK	-1.73304373871055	1.20480055890301e-08
ompF	1.36902711797722	1.24569984706603e-08
yidA	-1.81678123811936	5.27373817278108e-08
nupG	-1.68385950643439	1.51839338266006e-07
cydA	1.21898657234768	3.54023524452252e-07
ydjN	2.28898171915388	4.7729133067438e-07
gatZ	-1.6631435918259	6.41936103919016e-07
osmY	-2.19790739239054	1.32933907584465e-06
gatC	-1.8055075323	1.74509744013103e-06
icd	-1.03786035495016	1.78329990934835e-06
leuD	-2.47238776541326	2.25358518107105e-06
ahpC	-0.970478463015697	4.61200152666399e-06
aceE	0.862041733327824	8.99262173965245e-06
deoB	-1.11975017428218	9.15905330414087e-06
plaP	1.05638020467497	2.31225554070362e-05
fadL	1.0922137909467	2.31225554070362e-05
dnaK	0.875639928897247	3.02060879384669e-05
cspA	1.27673155628296	3.78956179238374e-05
ahpF	-0.964153733394864	9.901627510662e-05
gtrS	-1.1182075729314	0.000100105754464533
sdhA	-1.45595619343951	0.000105145798497739

gatA	-1.63891081142238	0.000106379913266369
acnB	-1.10470988654074	0.000123715275029271
ppsA	-1.74552496392493	0.000148624395383318
deaD	0.953538014485646	0.000226496881197669
pdhR	0.937777419625075	0.000296239671667935
sthA	-1.39506646196629	0.000406746473123539
aceF	0.764708695974235	0.000568385340805608
tsgA	1.14163901498538	0.000683668918995821
cydB	1.05613642749064	0.000811970359951299
nuoG	-1.09242101606278	0.0010126543333041
gpmA	-0.78533796955656	0.0023287411010205
poxB	-1.807134937772	0.00267030878297739
gatD	-1.48711732003352	0.00324866069920715
ndh	1.22193368114597	0.00327082256477578
fumA	-1.11641730135098	9.8339498118368e-05

2. ¿Qué genes se sobreexpresan para la solución de azúcar frente al E. Coli salvaje?

Se sobreexpresan los genes con valor positivo para la columna de log2(FC):

ptsG	3.70386159302384
ackA	1.34270180231898
ibpA	2.34243447457831
ompF	1.36902711797722
cydA	1.21898657234768
ydjN	2.28898171915388
aceE	0.862041733327824
plaP	1.05638020467497
fadL	1.0922137909467
dnaK	0.875639928897247
cspA	1.27673155628296
deaD	0.953538014485646
pdhR	0.937777419625075
aceF	0.764708695974235
tsgA	1.14163901498538
cydB	1.05613642749064
ndh	1.22193368114597

3. ¿Qué genes se infraexpresan para la solución de azúcar frente al E. Coli salvaje?

Se infraexpresan los genes con valor negativo para log2(FC):

setA	-3.77876481042538
deoC	-2.26082669697138
sucB	-2.04568679574465
sucD	-2.18476604460764
sodA	-1.53279491780711
deoA	-2.23164899075234
sucC	-2.19273413482056
sucA	-1.8763483981581
ydbK	-1.73304373871055
yidA	-1.81678123811936
nupG	-1.68385950643439
gatZ	-1.6631435918259
osmY	-2.19790739239054
gatC	-1.8055075323
icd	-1.03786035495016
leuD	-2.47238776541326
ahpC	-0.970478463015697
deoB	-1.11975017428218
ahpF	-0.964153733394864
gtrS	-1.1182075729314
sdhA	-1.45595619343951
gatA	-1.63891081142238
acnB	-1.10470988654074
ppsA	-1.74552496392493
sthA	-1.39506646196629
nuoG	-1.09242101606278
gpmA	-0.78533796955656
poxB	-1.807134937772
gatD	-1.48711732003352
fumA	-1.11641730135098

Enlace al historial de Galaxy: ***https://usegalaxy.org/u/bustos/h/prctica-genmica***