

Expansion and Standardization of Rat Expression Data at the Rat Genome Database

Wendy M. Demos^{1*}, Kent C. Brodie², Jeffrey L. De Pons¹, Adam C. Gibson¹, Varun Reddy Gollapally¹, G. Thomas Hayman¹, Mary L. Kaldunski¹, Akhilanand Kundurthi¹, Logan Lamers¹, Stanley J.F. Laulederkind¹, Jennifer R. Smith¹, Jyothi Thota¹, Marek A. Tutaj¹, Monika Tutaj¹, Mahima Vedi¹, Shur-Jen Wang¹, Stacy M. Zacher³, Melinda R. Dwinell¹, Anne E. Kwitek¹

Background

GEO

with beta-mercapto ethanol to preserve RNA integrity. Total RNA was isolated

using Rneasy micro kit as per manufacturer's suggestions (Qiagen), including

DNAse treatment. Total RNA was eluted in molecular grade water. RNAs with

SMART-seg2 was performed on total RNA as described by Picelli et al. Briefly

10 ng and 100 ng total RNA was used for RT. Resulting cDNA was pre-amplified

for 12 and 10 cycles, respectively and tagmentation performed on 18 ng and

20 ng cDNA using the Nextera DNA Sample Preparation Kit (Illumina).

ragments were amplified for 8 continuous cycles. Libraries sequenced as

L) Gene expression data are viewable from any Gene

optimal RIN scores were considered for library preparation

Rat Genome Database

Rat Genome Database, Dept. of Physiology¹; Clinical and Translational Science Institute²; Finance and Administration³ Medical College of Wisconsin, Milwaukee, WI 53226 USA

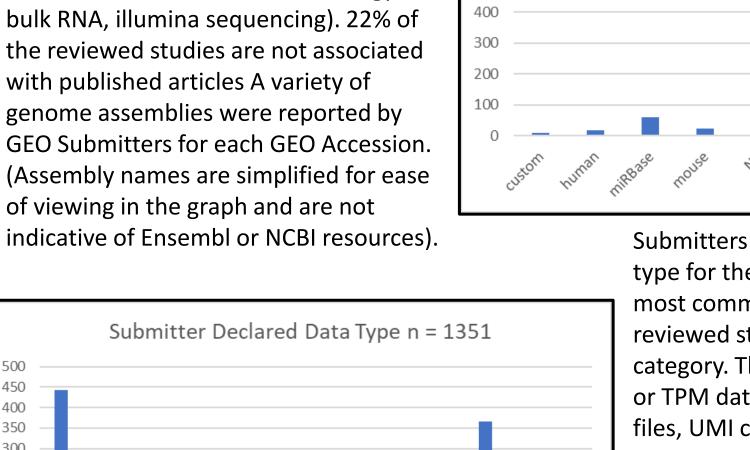
Abstract The Rat Genome Database (RGD) is expanding and incorporating expression data content into the larger ecosystem of RGD so users can seamlessly query for coherent gene information across portals. Researchers will be able to access expression metadata and values that were submitted to public resources such as the Gene Expression Omnibus (GEO) repository, with all expression values converted to transcripts per million (TPM). In Phase One of the project, an expression curation tool was developed to aid in comprehensive Natural Language Processing (NLP) assisted manual curation of

public datasets. The expression curation tool relies on a pipeline that imports metadata from the GEO Accession Display and utilizes NLP to match ontology terms to GEO series attributes. Curators can enter missing terms, confirm the predicted term, or provide a more specific term when appropriate. Fields for descriptors such as tissue type, vertebrate trait, clinical measurement, strain, cell type, experimental condition, etc. are built into the user interface. To enhance operational efficiency, the curation process begins at the project level interface, where ontology terms are entered a single time and then propagated across all applicable samples. Curators conduct a sample level review and edit terms on a per-sample basis as needed. When the metadata are correct and as complete as possible, they are loaded into the appropriate tables in RGD's relational database. The expression values submitted for the curated GEO series will be loaded for all genes and transcripts in the corresponding files. Currently, RGD has imported 1,859 GEO series related to Rattus norvegicus expression studies. Of those, 1,351 have been reviewed and prioritized for curation. To date, metadata for 165 GEO series have been uploaded. Expression value types submitted to the repository represent a wide range of analysis outputs (i.e., FPKM, counts, log₂FC). The submitter declared genome assemblies in the reviewed GEO series include versions RGSC3.4 -mRatBN7.2 as well as custom and non-rat references. The lack of standardization in the repository makes it difficult to identify rat data and furthermore, correlate expression values across studies. The goal of Phase Two is to standardize the expression values by developing and evaluating a bioinformatic pipeline that downloads and converts fastq files from the Sequence Read Archive, aligns to the most current and correct R. norvegicus genome assembly, and outputs TPM. This pipeline integrates quality control measures, alignment with the STAR¹ aligner, and abundance estimation with the RSEM² software package. Phase Three will focus on

enhanced visualization of expression values. The current tabular-based view will be updated and new

graphical visualizations at the gene and transcript levels are planned.

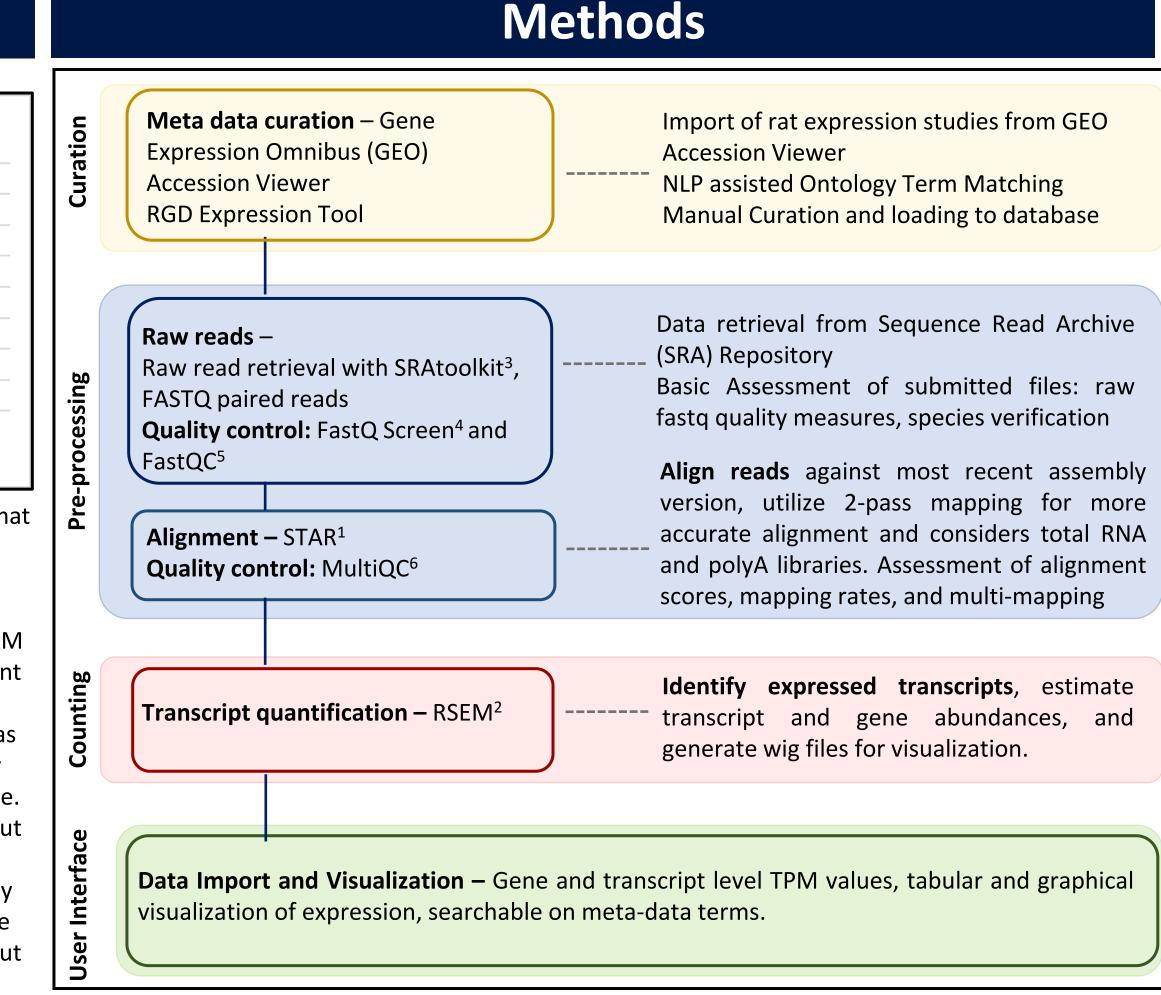
1,351 studies were reviewed at the time of writing the abstract for prioritization of meta-data curation. 58% are tagged for future curation due to stringent first pass curation criteria (data type = TPM, FPKM, source = non-culture, strategy = bulk RNA, illumina sequencing). 22% of the reviewed studies are not associated with published articles A variety of genome assemblies were reported by GEO Submitters for each GEO Accession. (Assembly names are simplified for ease of viewing in the graph and are not



Phase 1: Metadata Curation

Submitter Declared Genome Assembly n=1351 Submitters to GEO declare the data format

type for the provided results files. The most common data type in the 1,351 reviewed studies fall into the Other category. The *Other* category is non-FPKM or TPM data such as peak data, alignment files, UMI counts, etc. SuperSeries category may have multiple data types as a SuperSeries accession may have many SubSeries each with a different data type. At times, a data type will be provided, but upon review, the submitted data do not reflect the type, or there is a discrepancy between the filename and declared type (For example RPKM data are provided but the filename is ACCESSION.fpkm.txt).



S NCBI

Sample type

Characteristics

NCBI > GEO > Accession Display 2

Public on Jan 03, 2018

tissue compartment: dermis

Stage: Scab detachment day

skin status: Skin wound

Rat_dermis_rep3

GEO Public on Jan 03, 2018 and distinct transcriptome profile of epidermis Experiment type Expression profiling by high throughput sequencing We report whole tissue transcriptomes from rat and mouse wounds, as well as Total RNA was isolated from adult rat and mouse whole tissues. Tissues included rat and mouse wound dermis and wound epidermis collected one day after scab detachment, the time point that coincides with hair follicle Guerrero-Juarez CF, Astrowski AA, Murad R, Dang CT et al. Wound Induction of fat regeneration in The Regents of the Maksim V AR067273 skin wounds by hair follicle University of signaling California Last update date May 15, 2019 Contact name Maksim V Plikus 949-824-1260 Organization name University of California, Irvine Developmental and Cell Biology GPL20084 Illumina NextSeg 500 (Rattus norvegicus SM2550543 Mouse_dermis_rep3 49 Rat_dermis_rep: 1 Rat_epidermis_re 50553 Rat_epidermis_re 0555 Rat_IFE_rep2 556 Rat_IFE_rep3 0557 Rat_IFE_rep4 550558 Rat_IFE_rep5

MINIML 2

GSE97047_Rat_TPM.txt.gz

A) GEO Accession Display – Landing page for a given Geo Series. Experiment summary, submitter contact information, and publication ID (if available) are provided. Links to the individual sample detail web pages are provided (Boxed in blue.). Near the bottom of the page, download links for the entire series may be available, or data may be provided on a per-sample basis (https://www.ncbi.nlm.nih.g ov/geo/query/).

B) RGD Expression Curation tool project level curation page Where the NLP based prediction software finds a match, the original text from the GEO record that was matched as wel as the suggested ontology term are shown in the curation tool Curators can utilize the built-in ontology browsers to enter

missing terms, confirm the predicted terms or provide a more specific term when appropriate. The prepopulated data are verified by curators and missing information is manually gathered from the GEO Accession Viewer (highlighted in yellow in Figure A) as well as the publication and its references when available. If necessary, curators will reach out to corresponding authors to obtain clarifying information. Terms are entered a single time for a GEO Accession and propagated across all applicable samples. If a submitter provides more than 1 citation, additional PMIDs can be attached to the data.

B Geo Accession Id morphogenetic potential and C) GEO Accession Display – Sample RGD Pub Med IDs Level Details. Tissue Id: UBERON:0001003 Ont Tree Data found on individual sample page scan be Clinical Measuremen Ont Tree epidermis ribonucleic acid level select epidermal superficial used for projectselect epidermis gland + 66
select epidermis of feather
follicle + 66 Strain Id: level ((B) and cellLine ld: No cell lines imported individual samplecellType Id level curation. Age (in days) Low embryonic neonatal weanling 🗌 juvenile 🔽 adult 🔲 aged **Public Notes:** XCO:0000168 surgical manipulation No wound dressing was applied single full thickness excisional wound XCO:0000099

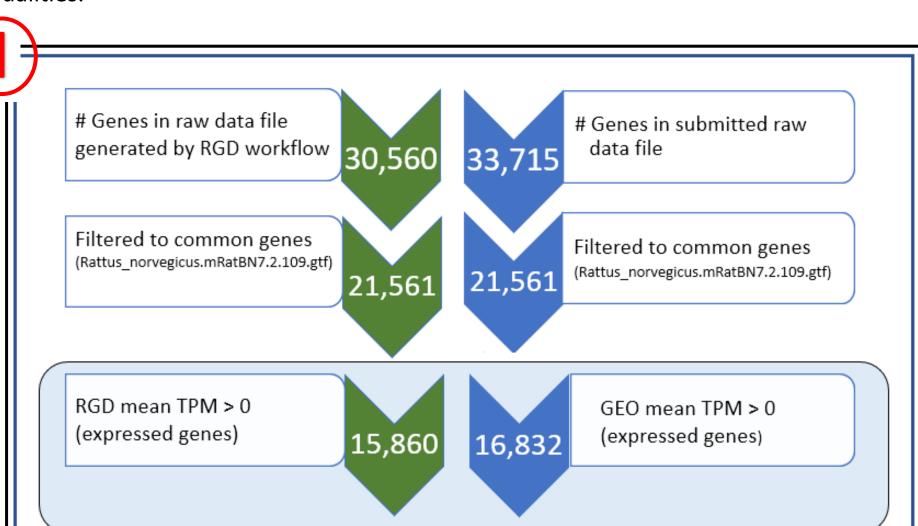
D) Individual sample details can be refined. Highlighted in red, brown, and purple are specific terms and conditions related to samples GSM2550549, GSM2550550, and GSM2550555 as displayed on the GEO Accession display (C). This differs from what was entered on the project-level page in the curation tool. Once the curation is complete, the

Curator will select the "Load Samples" button in the upper right corner and meta-data are loaded to RGD's relational

Phase 2: Data Re-analysis

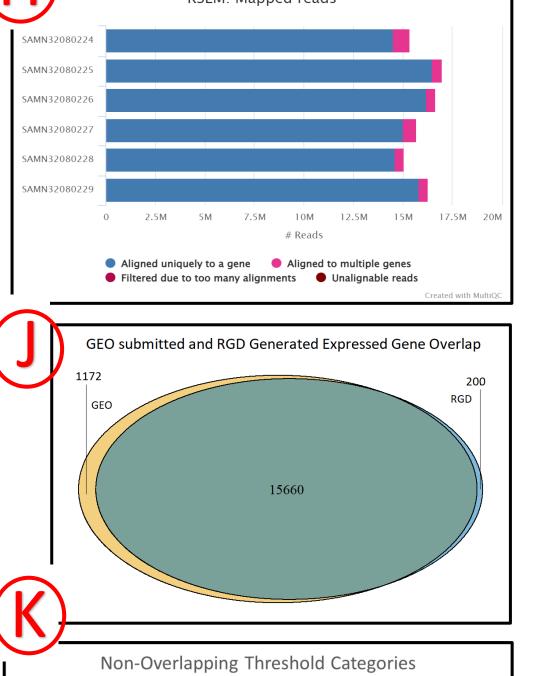
■ Multiple hits / multiple genom Multiple hits / one genome

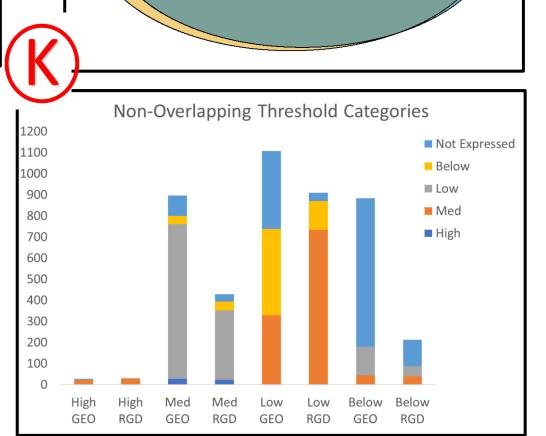
E) The Pre-processing stage of the data re-analysis pipeline utilizes FastQ Screen⁴ to verify the organism matches what is provided in the GEO Accession Viewer. (Inset) Both R. norvegicus and R. rattus assemblies are included in the tool to verify rat species. description and F) FastQC⁵ assesses the general quality of the SRA downloaded and converted FASTQ files (Data for all samples in series compiled by MultiQC⁶). MultiQC⁶ is used to assess the G) STAR² alignment and H) RSEM¹ mapping qualities.



I) Data for Accession GSE220261⁷ was used as a test to compare output by the RNAseq re-analysis pipeline. FPKM and count data were submitted to GEO and there is no associated publication. Other than knowing the data were aligned to "rn7", very few details were provided about the data. FPKM data were converted to TPM with R library GeoTcgaData⁸. Data were normalized by limiting the analysis to gene definitions in the Ensembl v109 gtf. J) The overall overlap of expressed genes between the submitted and RGD pipeline results was almost 92% for both the FPKM and TPM analysis. K) Using the expression thresholds as defined by the Expression Atlas Scale⁹ the genes categorized as medium expression had the highest overlap rate (85%). The graph shows the count and classification for non-overlapping expression results within each threshold category.

H RSEM: Mapped reads





Phase 3: Data Visualization

GEO

Public on Jan 03, 2018

Public on Jan 03, 2018

Rat inter-follicular epidermis

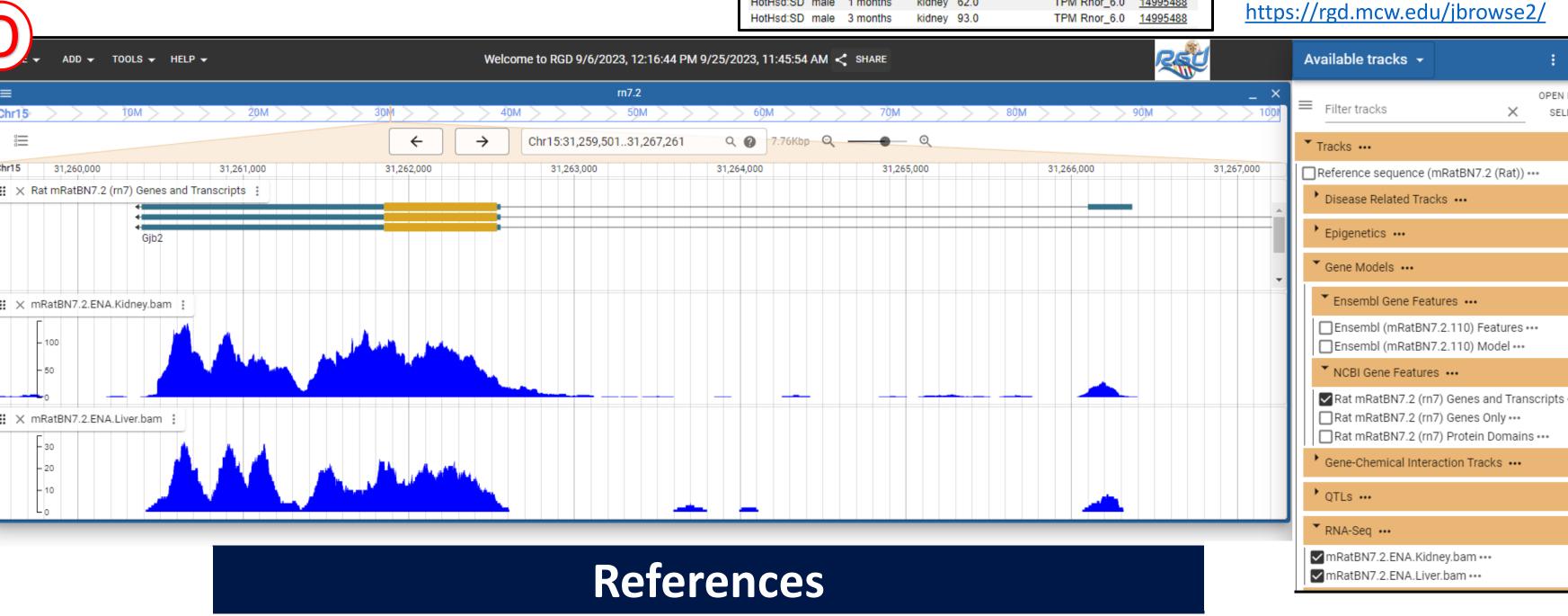
xtraction protocol Whole tissue was homogenized using Precellys in RLT buffer supplemented

with beta-mercapto ethanol to preserve RNA integrity. Total RNA was isolate

SMART-seg2 was performed on total RNA as described by Picelli et al. Briefly

ong and 100 ng total RNA was used for RT. Resulting cDNA was pre-amp

page on the RGD website (https://rgd.mcw.edu) Navigate to the expression data by using the link on the left-hand panel of the gene page. RGD Manual Diseas M) Users are directed to the RNA-SEQ Expression table which is organized by xpression Data Report for gene Gib2 system and the Expression Atlas Scale⁹ Rnor SHR UTH Rnor SHR Utx thresholds. N) Clicking on the View TPM Rnor_6.0 <u>13506920</u> RNA-SEQ expression Data or the TPM Rnor_6.0 <u>1350692</u> TPM Rnor_6.0 1350692 hyperlinks in the table brings users to TPM Rnor_6.0 <u>1350692</u> a detailed downloadable report. This TPM Rnor_6.0 <u>1350692</u> Sequence report includes specific sample TPM Rnor_6.0 <u>13506920</u> Nucleotide Sequence TPM Rnor_6.0 <u>14995488</u> information and links to the origin of TPM Rnor_6.0 14995488 TPM Rnor_6.0 1499548 the expression data and any TPM Rnor_6.0 14995488 Protein Structures TPM Rnor_6.0 <u>1499548</u>3 associated publications. O) Expression Additional Informat data tracks are viewable in RGD's TPM Rnor_6.0 1499548 Jbrowse2 instance TPM Rnor_6.0 <u>14995488</u> TPM Rnor_6.0 <u>14995488</u> TPM Rnor_6.0 14995488 Welcome to RGD 9/6/2023, 12:16:44 PM 9/25/2023, 11:45:54 AM < SHARE



- 1. Dobin A, Davis C, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras 5. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. T: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2012, Jan:29(1): 15-21
- 2. Li B, Dewey C: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011 (12), 323
- 3. SRAtoolkit SRA Toolkit Development Team Available from: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
- 4. Wingett SW and Andrews S. FastQ Screen: A tool for multi-genome mapping and quality 8. Hu E, GeoTcgaData. Available from https://github.com/YuLab-SMU/GeoTcgaData control [version 2; referees: 4 approved]. F1000Research 2018, 7:1338
- Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc
- 6. Ewels P, Magnusson M, Lundin S, Käller M: MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016)
- 7. Li X: **GSE220261** Available from:

https://www.ebi.ac.uk/gxa/help/index.html

- https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220261
- 9. EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.
- This and other recent RGD presentations are freely available for viewing









