23Q4 Release Notes

Overview of the 23Q4 Release

We're excited to share with you several new genomic and CRISPR data and pipeline improvements in this release!

Read on to learn more about new data and changes.

Metadata

Metadata files

New in 23Q4 and beyond, each release dataset will have a downloadable data dictionary that defines and describes all available metadata columns.

A reference column (info_reference) indicates whether a specific column is referencing the model or the patient from which the model was derived.

New screens

In this release, you'll find 28 new CRISPR screens across the following lineages:

- Biliary Tract (n=1)
- Bowel (n=1)
- Breast (n=4)
- CNS/Brain (n=1)
- Esophagus/Stomach (n=1)
- Eye (n=1)
- Head and Neck (n=2)
- Kidney (n=1)
- Lung (n=4)
- Lymphoid (n=5)
- Ovary/Fallopian Tube (n=3)
- Pancreas (n=1)
- Skin (n=2)
- Vulva/Vagina (n=1)

New Datasets

New Humagne Library Screens

In this release, you will notice that we have begun conducting CRISPR screens in a new library: Humagne (version CD) from the Genomics Perturbation Platform at the Broad Institute. Humagne is a dual-knockout Cas12 library with two constructs per gene and two sgRNAs (targeting the same gene) per construct. For more details, read here

The smaller size of the library and improved on-target efficacy from having two sgRNAs per construct will enable us to screen more difficult models.

Cell models with Humagne libraries are generally screened at 14 days. Please see the CRISPR metadata for more details about the duration of the screens.

Omics Pipeline Updates

Mutation calling pipeline updates

We've updated our mutation calling pipeline to include more recent annotations, better germline filtering and more comprehensive rescuing of cancer-related variants. Please read more about the updates to our mutation calling pipeline here.

Cell Lines with Legacy Copy Number have been dropped from the dataset

Legacy Copy Number refers to a mixture of SNParray and WES profiles, for which we don't have raw sequencing files. In 22Q4, we removed all "Legacy Copy Number" data from the Mutations file.

This has impacted our ability to release some previously passing CRISPR screens in 23Q4. We will re-release the screens once new Copy Number data has been generated.

Copy Number and Mutation: Known segmental duplication regions

Due to the limitation of short-read sequencing, regions that are duplicated on the genome typically have inaccurate calls. Therefore, for gene-level copy number and mutation data, we are masking genes that overlap with known segmental duplication regions and/or are flagged by repeatMasker. For details, please see our github.

New Omics Data

New Profile-level Expression data

We are now releasing 2 additional Profile-level expression files that enable more flexible normalization and interrogation of gene expression data:

- OmicsExpressionAllGenesTPMLogp1Profile: Contains normalized TPM values for *all* genes
- OmicsExpressionAllGenesEffectiveLengthProfile: Contains the effective gene lengths output by RSEM.

CRISPR Pipeline Updates

Removal of Confounding Components (RCPC) for Screen Quality Correction

We've found that screen quality explains several components of variance in the gene effect matrices outputted by Chronos. To remove this effect, we applied a correction called the Removal of Confounding Components (RCPC). We observed that doing so improved the data quality by all metrics (including NNMD, ROCAUC, and FPR).

First, PCA was performed on ScreenGeneEffect and CRISPRGeneEffect with column variance normalization, indexed by screen/cell line. Missing values were imputed using the column mean before performing PCA. The component loadings were then correlated with the median depletion of common essential genes. Component loadings that had a Pearson correlation greater than 0.2 were removed. The truncated component loadings and principal components were then multiplied to revert back to the matrix of original dimensions. Column variance was rescaled to match the column variance of the original matrix. Finally, the mean imputed missing values were removed.

More information on RCPC can be found here.

Legacy Copy Number has led to new failing Screens

Dropping legacy copy number data has led to the following screens now failing our data QC:

- SC-000014.AV01
- SC-000600.AV01
- SC-000658.AV01

- SC-000854.AV01
- SC-000282.KY01
- SC-002217.KY01

Once new copy number data is available, these screens will be re-released.

Portal Tools

Celligner Update

In this release, you will notice significant updates to the Celligner tool!

Celligner has been more closely integrated into our pipelines and as such will now be continually updated with new cell lines and models in each release.

Further, we’ve updated the interface to allow highlighting cell lines and models by context (such as mutation identity, RNA expression, etc.), selecting cell lines to create new contexts that can be analyzed in Data Explorer and coloring models by growth pattern (adherent, suspension, etc).