Archer® Analysis 6.0 User Manual

For software version(s): 6.0

Released 2019-03-22

Notices

Limitations of Use

For Research Use Only. Not intended to be used for diagnosis or treatment of human or animal diseases. This product was developed, manufactured, and sold for research use only.

Symbols Used In This Manual

Attention symbols denote critical information about additional resources needed, key workflow steps, limitations of use, or where risk of assay failure could occur if not carefully observed.

Reminder symbols call attention to minor details that may be easily overlooked and compromise the procedure resulting in decreased assay performance.

In This User Manual

This manual is a guide to using Archer Analysis, a comprehensive bioinformatics suite designed to be used in conjunction with Archer library preparation kits and gene enrichment panels for targeted sequencing of select genes and regions of interest using next-generation sequencing (NGS).

Overview – Page 3

This section contains the intended use statement, test principle, an overview of the scope of the software, as well as a description of how the software can be accessed.

Before Getting Started – Page 4

This section contains critical guidance for the successful implementation of the functions and workflows described in this manual, including additional software that may be required, but not supplied by Archer. **This should be read and understood before data analysis is initiated.**

Detailed Software Features and Workflows – Page 6

This section provides detailed definitions and step-by-step instructions for the features and workflows available in the software. A table of contents is provided for quick reference to specific headings.

Additional Resources

View videos and additional resources for Archer products at:

https://archerdx.com/support

Technical Support

Visit https://archerdx.com/support/faqs for a list of helpful answers to frequently asked questions or contact us directly at tech@archerdx.com.

Overview

Intended Use

Archer Analysis is a comprehensive bioinformatics suite that is intended for use only with next-generation sequencing (NGS) data generated by Archer reagent kits and corresponding target-enrichment panels on either Ion Torrent™ or Illumina® platforms. Analysis provides functions for all steps of secondary analysis (read cleaning, alignment and mutation calling), as well as some steps for tertiary analysis (interpretation). The software can be used with any Archer panel, without need for installing additional modules or other special configuration.

Test Principle

AMP™, or Anchored Multiplex PCR, is a rapid and scalable method to generate target-enriched libraries for NGS. AMP technology can be used for applications in targeted RNA sequencing, genomic DNA sequencing and genotyping applications to generate a sequencing library in a matter of hours. Designed for low nucleic acid input, this process delivers robust performance across a variety of sample types.

AMP utilizes unidirectional gene-specific primers (GSPs) that enrich for both known and unknown mutations. Adapters that contain both molecular barcodes and sample indices permit quantitative multiplex data analysis, read deduplication and accurate mutation calling.

The Archer Analysis software utilizes molecular barcodes for error correction and read deduplication to support quantitative multiplex data analysis and confident mutation detection. Analysis reports both sequencing metrics and number of unique observations supporting called mutations.

How to Access the Software

Archer Analysis is a web-based software, using an Apache server and accessed via any modern browser, with the exception of Internet Explorer® version 7 (Windows® XP), which has been shown to have some problems. Google Chrome™ is the recommended web browser. A hosted version of the software (private to your organization) is available via Archer's Analysis Unlimited service. A free demo of this service can be accessed at https://analysis.archerdx.com/.

If local installation is required, your lab will need to secure the requisite computing resources (see the Before Getting Started section of this manual), as the Analysis server is packaged as a virtual machine (i.e., must run in a virtualization environment such as VMware ESXi®).

Please visit https://archerdx.com/software/analysis or contact tech@archerdx.com for further details and assistance in accessing any of the options listed above.

Before Getting Started

Resources Required To Operate Archer Analysis

While the user interacts with Analysis via a standard web browser, data is processed, stored, and served via the Analysis Server package. Thus, this server must be accessible to your web browser and running on one of these options:

1. your local machine
2. other hardware on your local intranet
3. a remotely hosted location on the internet

Archer provides option #3 as a private, paid service for your organization (including HIPAA compliance) via its Analysis Unlimited service, which automatically and dynamically allocates computing resources to process any number of samples submitted by the user.

If option #1 or #2 are chosen, then the Analysis server must be installed on a VMware based hypervisor. Of those we recommend ESXi for best performance but also support Workstation and Fusion.

Archer does not provide virtualization software to run the Analysis Server VM locally; this must be obtained separately by your organization. VMware is recommended.

The computing resources required to run Archer Analysis depend on your lab's required throughput, i.e., the number of samples and total processing time required per batch. The VM requires:

- at least 1TB of storage space
- one dedicated CPU core, for each sample to be processed in a batch
- 10 GB of RAM, for each sample to be processed in a batch.

For example, in order to process a 4-sample job, 4 dedicated processing cores and 40 GB of dedicated RAM are required. Please refer to the separate **Archer Analysis Virtual Machine Installation Guide** for detailed instructions on installation and system requirements of the VM.

Preparing Your Data For Archer Analysis

Samples are processed in Archer Analysis as a "job". One or multiple samples can be run together in a single job.

Note these factors when preparing your data for an Analysis job:

- All samples must have been prepped using the same Archer panel (but not necessarily at the same time).

- All samples in a job will be processed using the same Analysis pipelines and settings that are assigned at job setup.

There are three types of data files that Archer Analysis accepts as input when setting up a new job, two required and one optional. These files are explained in the sub-sections below:

Demultiplexed FASTQ (Illumina) or BAM (Ion) file(s) – REQUIRED

For each sample to be processed in a job, the corresponding FASTQ (for Illumina data) or BAM (for Ion data) file must be uploaded to Archer Analysis (both R1 and R2 files are required for samples that are paired-end sequenced).

Note these factors when preparing your FASTQ files for a job:

- Software to demultiplex FASTQ files is NOT provided by Archer, and is instead provided by the sequencing instrument vendor.
- Aside from demultiplexing, no other preprocessing of the sequencing data is required (Analysis performs all read trimming, cleaning, etc.)
- All sample files must be in the same folder and selected at the same time.

Target Regions file – REQUIRED

This is a GTF file that is unique to the Archer panel used for library preparation of the sample(s) to be processed in a job. It is supplied by Archer, and is usually already packaged in Analysis. Consult the product literature for the specific Archer panel to determine which file should be used. If you do not see the required file in Analysis, it can be obtained by contacting tech@archerdx.com, and uploaded after installation, as described in *Managing and Adding GTF Files*.

Targeted Mutations files – OPTIONAL

If there are specific mutations of interest when running the SNV/InDel pipeline (for either RNA or DNA panels), then a list of these can be compiled in Variant Call Format (VCF, v4.1) and supplied to the software before processing a batch of samples. This will force the software to report on each mutation in the list, regardless of being called by the pipeline.

Using a Targeted Mutations file with the SNP/InDel pipeline is an excellent way to quickly surface information on mutations of known significance for your application. It can be used while simultaneously searching for non-targeted variants, so that no stone is left unturned.

Detailed Software Features and Workflows

INTRODUCTION 12
PIPELINES INCLUDED IN ANALYSIS 12
FusionPlex® (RNA) Assays: 12
VariantPlex® (DNA) and Reveal ctDNA™ Assays: 12
Immunoverse™ (RNA) Assays: 12
PRIMARY NAVIGATIONAL FEATURES 13
Main Header & Main Menu 13
Dropdown Menus 13
Footer: Legal & About Analysis Pages 13
Info Pop-ups 14
OUTPUT OPTIONS, SYSTEM ARCHITECTURE AND ALGORITHM DETAILS 14
LOGGING INTO ANALYSIS 15
Creating an Account 16
The Home Page 17
Logging Out of Analysis 17
MANAGING USERS AND GROUPS 18
The Admin User 18
Users 19
Adding and/or Editing Users 19
Deleting Users 23
Groups 23
Adding and Editing Groups 24
Deleting Groups 26
MANAGING SYSTEM (GTF AND VCF) FILES 26
Managing and Adding GTF Files 27
GTF File Actions (View, Download & Visibility) 27
Deleting GTF Files 29
Managing VCF Files 29
USER SETTINGS (APPLIED ACROSS ALL JOBS) 30
Viewing and Changing Settings 30
General Analysis Settings 31
MIN_AVERAGE_UNIQUE_DNA_AND_AMBIG_START_SITES_PER_GSP2 31
MIN_AVERAGE_UNIQUE_RNA_START_SITES_PER_GSP2_CONTROLS 31

ERROR_CORRECTION 32
READ_DEPTH_NORMALIZATION 32
BARCODE_LENGTH 32
DEBUG 32
STRUCTURAL VARIATION ANALYSIS SETTINGS 32
DISPLAY_INTRONIC_STRUCTURAL_VARIATIONS 33
MIN_READS_FOR_VALID_STRUCTURAL_VARIATION 33
XCONTAM_SINGLE_PANEL 33
XCONTAM_FUSION_CONFIDENCE_THRESHOLD 34
EXPRESSION_IMBALANCE 34
MIN_UNIQUE_START_SITES_FOR_VALID_STRUCTURAL_VARIATION 34
STRUCTURAL_VARIATION_PERCENT_OF_GSP2_READS 34
EXPRESSION_IMBALANCE_ALPHA 34
CALL_TRANSCRIPTIONAL_READTHROUGH_EVENTS_WEAK 34
RNA_EXPRESSION_VISUALIZATION 34
RNA_EXPRESSION_OUTLIERS 34
SNP/INDEL PIPELINE SETTINGS 34
MIN_DEPTH_FOR_VARIANT_CALL 35
MAPQ_THRESHOLD_FOR_VARIANT_CALL 35
MIN_BASEQUAL_FOR_VARIANT_CALL 35
MIN_ALLELE_FRACTION_FOR_VARIANT_CALL 35
MIN_PHRED_QUAL_SCORE_FOR_VARIANT_CALL 36
VARIANT_DOWNSTREAM_ROI_SIZE 36
OUTLIER_DETECTION 36
MIN_DAO_FOR_OUTLIER_DETECTION_CALCULATION 36
DEEP_SHALLOW_THRESHOLD 36
CNV PIPELINE SETTINGS 36
CNV_STRONG_AMPLIFICATION_THRESHOLD 36
CNV_STRONG_DELETION_THRESHOLD 37
CNV_P_VALUE_THRESHOLD 37
MISCELLANEOUS SYSTEM SETTINGS 37
JOBS_PER_PAGE 37
LANDING_PAGE 37
CFDNA_READ_DEPTH_NORMALIZATION 37
SAVE NEW USER SETTINGS OR RESTORE DEFAULTS 37
IDENTIFY WHICH SETTINGS ARE NOT DEFAULT 38
JOB SETUP AND MANAGEMENT 38
STARTING A JOB: PERFORM ANALYSIS PAGE 38

CHOOSING THE PIPELINE(S) 41
RNA Analysis Types 41
DNA Analysis Types 42
Selecting Normalization Parameters for CNV 44
DNA SNP/InDel Detection Types 45
Target Region (GTF) File Annotations 45
ADVANCED SETTINGS 46
ANALYSIS TEMPLATES 47
RUNNING JOBS 47
Progress Bar and Status Updates 48
Analyzing Multiple Samples & Jobs 48
Stopping a Running or Queued Job 49
Jobs That Complete with Errors 50
Job Archival States 51
Unarchiving Jobs 52
JOB HOOKS 53
ADDING A JOB HOOK 53
MANAGING EXISTING HOOKS 56
Editing Existing Hooks 56
Disabling Existing Hooks 56
Enabling hooks on a per job basis 56
Deleting Existing Hooks 57
VALIDATION SETS 57
MARKING SAMPLES FOR VALIDATION SET 58
CREATING/REMOVING A VALIDATION SET 58
RUNNING A VALIDATION SET 60
VALIDATION REPORT 61
RUN HISTORY 62
ACCESSING AND INTERPRETING RESULTS 62
KEY ASSAY CONCEPTS 62
Genome Reference 62
Raw Reads, Unique Fragments, and Unique Start Sites 62
RNA vs. DNA vs. Ambiguous Reads 63
Sample Tracking 64
Control Genes (FusionPlex Only) 65
GENERAL RESULTS REVIEW & INTERPRETATION WORKFLOW 65

Batches 66
Recent Samples 66
Sample Summary 67
QC Results 68
Viewing and Overriding QC Pass/Fail Threshold 69
Results Review and Reporting 70
Past Results 70
Reviewing Settings of a Job 71
Reviewing (Error) Logs 72
Rerun or Clone a Job 73
View Overview of Results for the Entire Job (Reports) 73
Job Details Page (Sample Summary) 74
User Defined File Download List 75
Visualize Sample Reads in JBrowse 76
Sample Details Page 77
Read Statistics 77
Assay Targets 79
Pipeline-Specific Tabs 81
Pipeline-Specific Information 81
Fusion/Isoforms (RNA Only) 81
Lymphoma Cell of Origin (COO) Pipeline and Results (Experimental) 88
SNP/InDel (RNA and DNA) 90
Copy Number Variation (DNA Only) 93
Immune Repertoire (RNA Only) 94
Outlier Detection 101
NORMAL DATA SETS (SNP/INDEL & CNV) 103
Normal Data Sets Terminology 104
Performing Analysis to Create Normal Data Sets 105
User Settings for Normal Data Sets 105
Creating a Normal Data Set 105
Sample Visibility 106
Adding a Sample to an Normal Data Set from Job Details 107
Existing Normal Data Sets 109
Performing Analysis with Normal Data Sets 110
Watched Folders and Normal Data sets 111
Deprecated and Upgrading Normal Data Sets 112
Accessing Experimental Features 112

Experimental Features for Job Set Up 113
WORKFLOW AUTOMATION AND INTEGRATIONS 113
WATCHED FOLDERS 113
Set Up a Watched Folder 113
Analysis Settings for the Executed Workflows 114
Removing a Watched Folder 114
Removing Target Region Files That Are Used in Watched Folder Workflows 114
Develop Procedure/Script For the Movement of FASTQ or BAM Files 114
How the System Knows When to Execute a Workflow 115
Example for a Job Containing a Single Sample 115
Example for a Job Containing Multiple Samples 116
Fate of Files in the Watched Folder 116
File Collisions 117
S3 Watched Folders 118
SSL-Enabled Linked Analysis Servers 118
TROUBLESHOOTING 119
JOB FINISHED WITH JOB STATUS "COMPLETED_ERROR" 119
JOB IS STUCK WITH JOB STATUS "NEW" 120
JOB IS STUCK WITH JOB STATUS "HOLD" 121
APPENDIX 121
GLOSSARY OF GENERAL TERMS USED IN THIS MANUAL 121
FORMAT OF THE GTF FILES 121
FORMAT OF THE TARGETED MUTATION FILE (TMF) 122
READ STATS & QC METRICS 123
Overall Sample Read Statistics 123
Fusion/Isoform Metrics 124
OUTPUT FROM 'USER DEFINED FILE DOWNLOAD LIST' 124
Sample-Level Data 124
Job-Level Data 125
Full Results File (full_results.txt) 125
BAM Files 132
VCF Files in "VCFs" Subfolder 133
VCF Files in "summaries" Subfolder 134
CNV Files 134
RNA Expression Files 135
Immune Repertoire Files 135

VARIANT SUMMARY GRID COLUMNS ____ *135*
General Variant Info ____ 135
Local Variant History ____ 139
Advanced Statistical Metrics ____ 140
Biological Significance ____ 140
SUMMARY OF ICONS ____ *146*
TECHNOLOGY LICENSES ____ 148

Introduction

Archer Analysis is comprehensive bioinformatics suite that includes pipelines for all of Archer's targeted NGS assays, both catalog and custom. The software covers all of secondary analysis (e.g. read trimming, QC, alignment, and mutation calling), as well as initial steps of tertiary analysis (e.g. annotations from Ensembl Variant Effect Predictor (VEP) and ClinVar):

- http://uswest.ensembl.org/info/docs/tools/vep/index.html
- https://www.ncbi.nlm.nih.gov/clinvar/

Although Analysis comes configured with carefully chosen criteria for mutation calling, it is important to fully understand all features of the pipeline(s) being used, such that settings can be optimized to meet the specific requirements of your lab and experiments. Bioinformatics specialists (tech@archerdx.com) are available to assist in this process. Once complete, Analysis is an excellent tool for routine use by both expert and novice users.

Pipelines Included in Analysis

The pipelines available in the software are as follows (pipelines within a given product group can be run in parallel):

FusionPlex® (RNA) Assays:

- **Fusions/Isoforms** – to call fusions and other structural rearrangements
- **SNP/InDel** – to call point mutations and insertions/deletions up to ~30bp in length
- **Expression** – for determining relative expression levels and imbalances
- **Lymphoma Cell of Origin (COO)** – experimental feature to identify the cell of origin subtype from a diffuse large B-cell lymphoma (DLBCL)

VariantPlex® (DNA) and Reveal ctDNA™ Assays:

- **Structural Variations** – to call exon skipping, internal tandem duplications (ITDs), and other structural rearrangements.
- **SNP/InDel** – to call point mutations and insertions/deletions up to ~30bp in length
- **CNV** – to identify relative copy number gains/losses at both the primer and gene-level

Immunoverse™ (RNA) Assays:

- **Immune Repertoire** – to identify total clonotypes (diversity), specific clonal frequencies, and hypermutational status for TCR and BCR populations

Primary Navigational Features

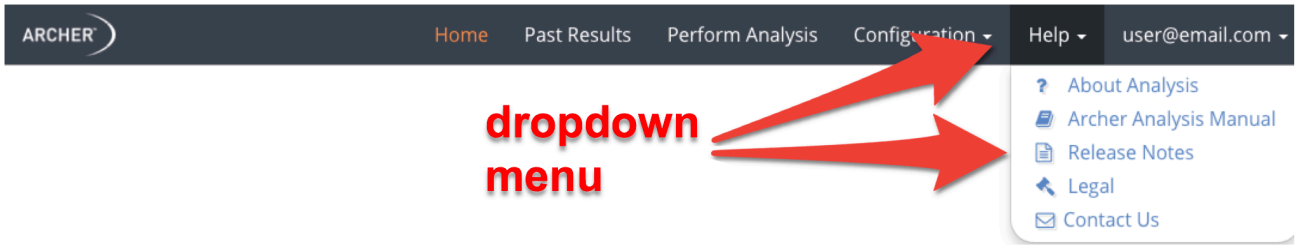
Archer Analysis features the same core navigational elements to ensure a consistent and intuitive experience across the entire system. Additionally, once logged in, every page has the same core structural elements, containing either links to key parts of the system and/or informational elements intended to guide use of the system. These navigational and structural elements are listed below.

Main Header & Main Menu

The top of every page features a header with the same **Main Menu** items:

Dropdown Menus

Dropdown menus are used throughout the system. Below is the **Help** option from the Main Menu:



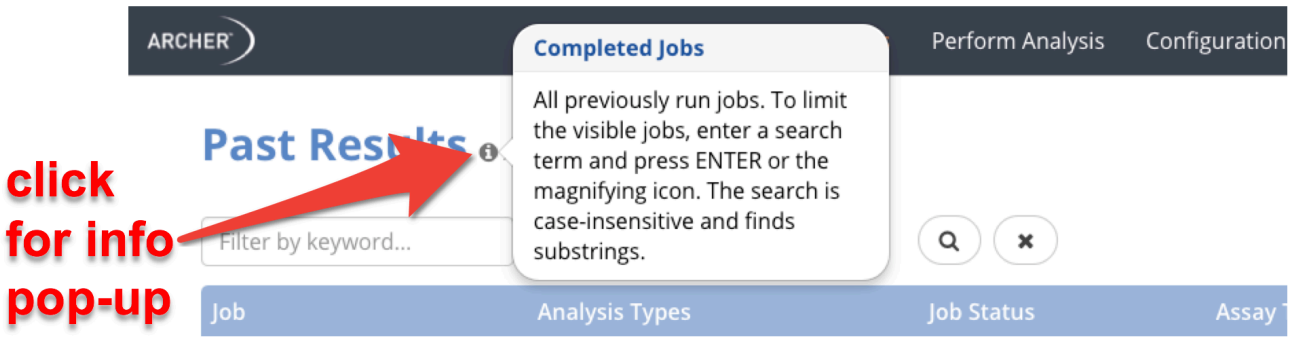
Footer: Legal & About Analysis Pages

The bottom of every page features a footer that includes links to the Legal and About Analysis pages, which detail information on the exact version of Archer Analysis, which third-party components are used, etc.:

Curious which version of Analysis you're running? Check out the **About Analysis** page. Also, make use of the icon that will provide information pop-ups.

Info Pop-ups

Whenever an icon appears next to terms and metrics throughout Analysis, the user can click or hover over the symbol to see descriptive information pop-up. Make use of these for quick reference on definitions and explanations of how such parameters are being used in the software:



Output Options, System Architecture and Algorithm Details

Most of the results from Analysis can be viewed in its graphical user interface (GUI), which is accessed via a standard web browser. Full results, including processing logs, can also be accessed as a combination of flat text, BAM, image, and PDF files, as well as via a REST API. This manual focuses on the GUI but see *Reports and Other Output Files* and *Output From Download All Files* for additional information regarding these features.

Analysis is built using several open-source, well-documented and stable technologies, namely CentOS (Linux operating system), Apache (web server) and PostgreSQL (database server). The bioinformatics pipelines listed above also make use of third-party modules. A comprehensive list of these modules can be found by clicking the "About Analysis" link located in the footer section of any page in the GUI:

Note, this manual is intended primarily to guide routine use of the software, and **does not** provide an exhaustive explanation of the methods and algorithms that underpin the various pipelines and functions available. If needed for security, privacy, or other concerns of your lab, additional documentation is available on the below topics and can be obtained by contacting tech@archerdx.com:

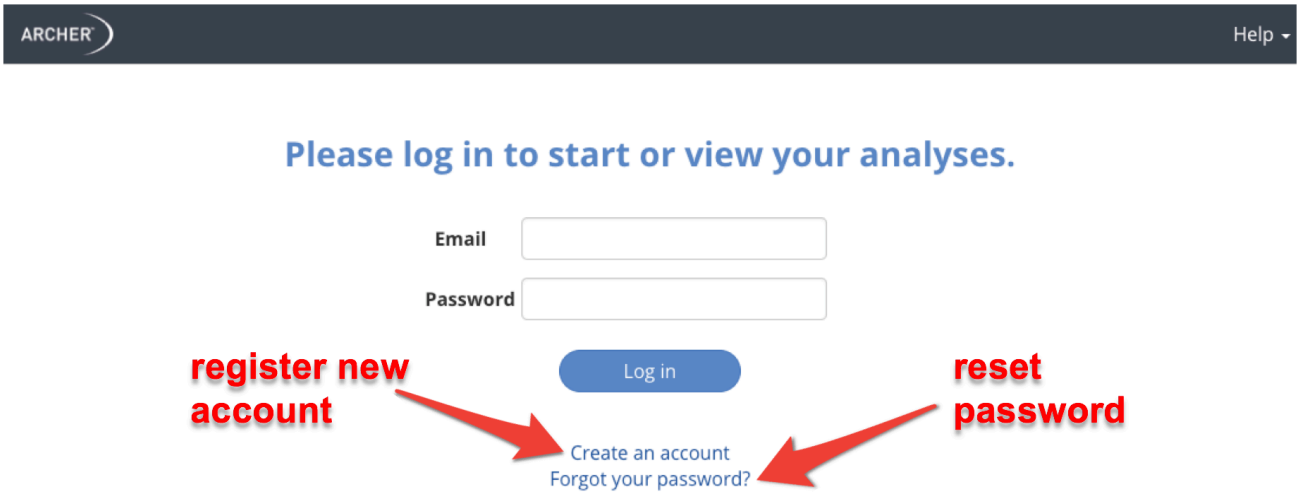
- Installing the software locally as a VM
- General system architecture

- Details of each bioinformatics pipeline listed above in this section
- PDF Reporting API
- REST API
- Process Logs

Logging into Analysis

The Archer Analysis Software is accessible via most modern web browsers, with the exception of Internet Explorer® version 7 which is incompatible and not supported. Google Chrome™ is the recommended web browser. Refer to the **Archer Analysis Virtual Machine Installation Guide** for instructions on how to discover the IP address of the Archer Analysis software. Start a web browser application and enter the IP address for the virtual machine in the address box.

The Archer Analysis Software is a fully contained and secure environment that allows users to run their analyses under their own account. The Archer Analysis login screen will allow the user to sign in to their account, register for a new account, and reset their password.



If the login page does not appear, ensure that the virtual machine is running, and accessible by the IP address or URL designated by your site administrator (see the **Archer Analysis Virtual Machine Installation Guide** for instructions on how to install the virtual machine).

There is a **limit of 5 login attempts for incorrect passwords**, after which your account will be de-activated. You will then be required to supply answers to the two security questions you specified at account registration, in order to re-assign your password. Alternatively, an admin user can manually re-activate your account and assign a temporary password.

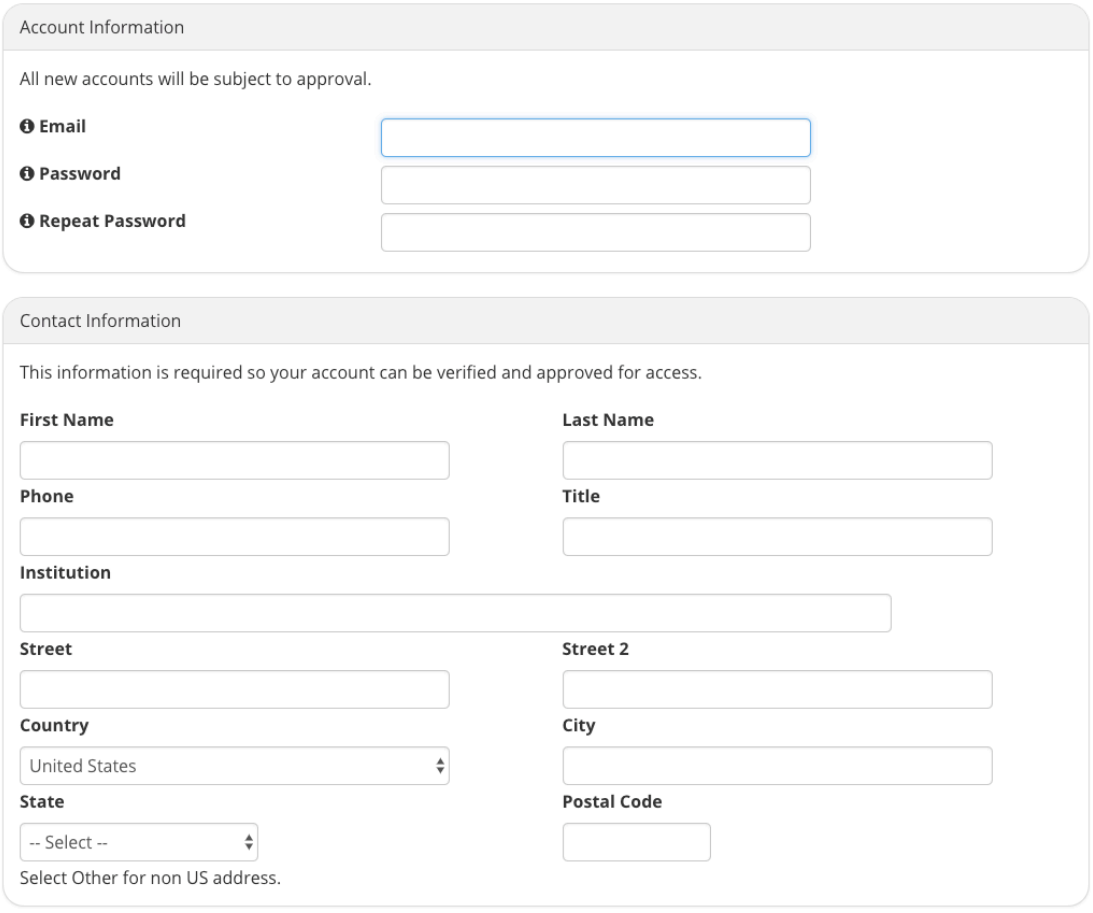
Creating an Account

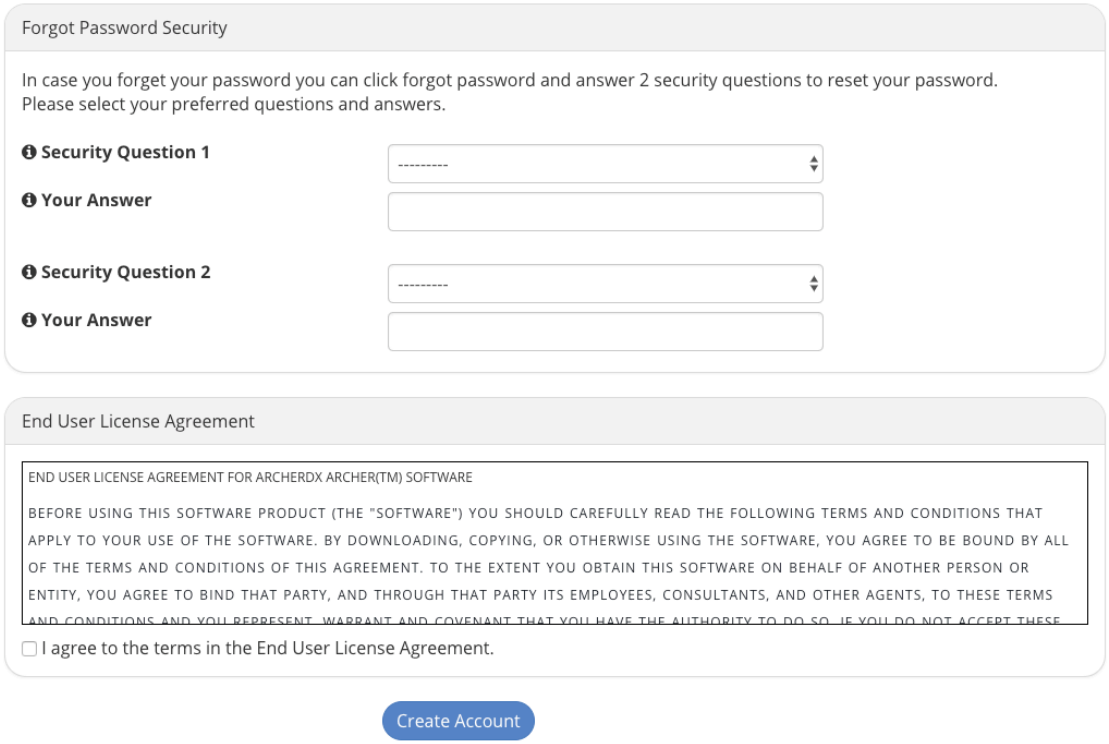
If the “Create an account” link is chosen, the account creation page will appear as shown below. Use your email address as your login, create a password, and specify two security questions/answers (to reset your password in case it is forgotten). For security reasons, passwords require the following complexity criteria:

- Your password can't be too similar to your other personal information.
- Your password must contain at least 9 characters.
- Your password can't be a commonly used password.
- Your password must contain 1 uppercase letter,1 lowercase letter,1 digit,1 special character.

Read the End User License Agreement (EULA) and check the box to indicate you have read and agree to the EULA, then click the “**Create Account**” button to create your account. This will automatically log you in.

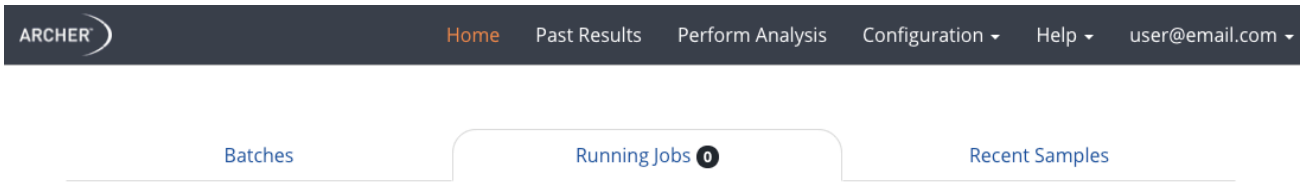
Alternatively, your system administrator can create a login for you (see Adding and/or Editing Users for more details), although you will still be required to follow the step above upon your first login:





The Home Page

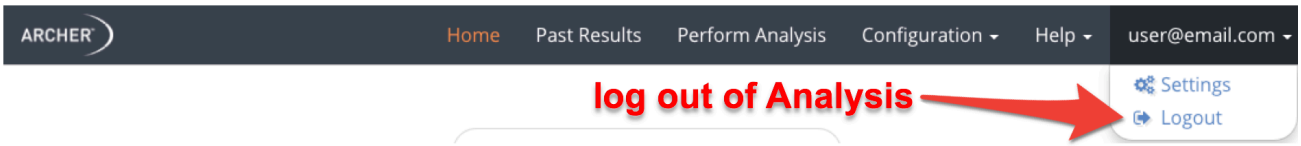
After successful login, the default home page will be displayed as shown below. Note, if you have been assigned admin privileges, or assigned to a User Group with jobs sharing enabled, you may see jobs already started by other users on this page:



The landing page is configurable within "**<user_email>**" → "**Settings**" → "**LANDING_PAGE**". User has the ability to select the designated landing page.

Logging Out of Analysis

You can log out of the system by accessing the dropdown menu next to the username displayed in the upper right corner of every page in the software:



For security purposes, Analysis will automatically log out any user (including admins) after 30 minutes of inactivity.

Managing Users and Groups

Archer Analysis utilizes user groups to control the visibility of job and sample data, as well as access to global system settings. There are two user types, admin and basic users, and additional permissions and data access granularity can be defined via custom user groups. Software permissions can be applied on a per-user and a per-group basis. Users can be a member of multiple groups, and permissions are additive for the user and any of their groups.

The Admin User

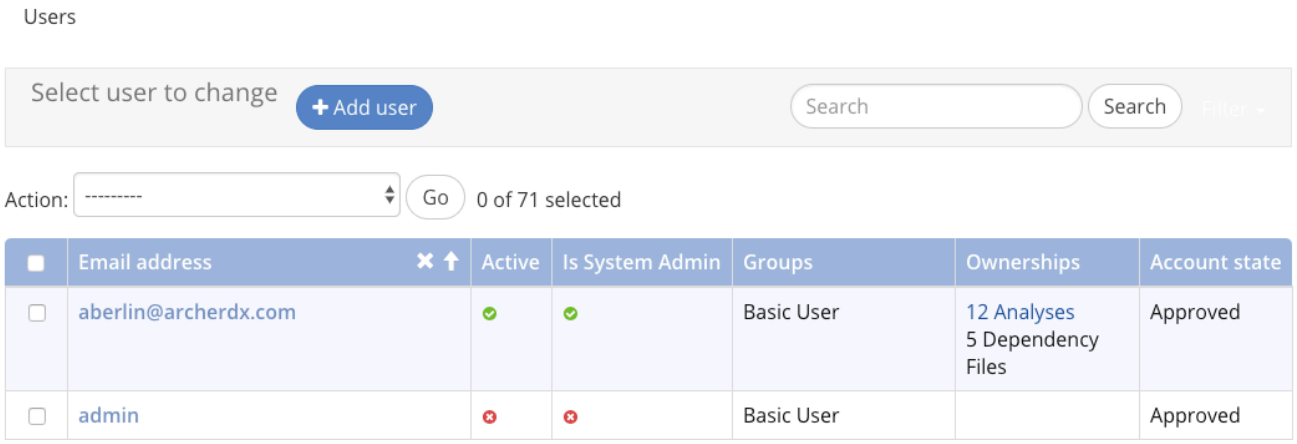
The admin user is by default a member of the "Basic Users" and "Experimental" groups. It also has a special permission called "Is System Admin", which is required in order to perform any of the user, group, and permissions management described in this section. This functionality is accessed under the **"Admin"** option in the Main Menu, by selecting either **"Users"** or **"Groups"**:

Note, Archer Analysis **does not** come configured with a default admin user. Therefore, one of the primary steps in installing Analysis is to set up the first admin account via the command line interface (CLI). Please refer to the **Archer Analysis Virtual Machine Installation Guide** for further details.

Admin users automatically have several privileges over standard (“basic”) users. This includes: (1) user and group management, (2) access to experimental features, (3) ability to view all jobs in the system, (4) ability to change ownership of jobs to any user, and (5) management of Targeted Mutations (VCF) and Assay Targets (GTF) files.

Users

Selecting the **"Admin"**>**"Users"** menu item takes you to a table containing a list of all the users in the system. The columns indicate the status of each user, whether they are a designated system admin, the groups they belong to, how many jobs and dependency files they have ownership of and the approval status of the account.

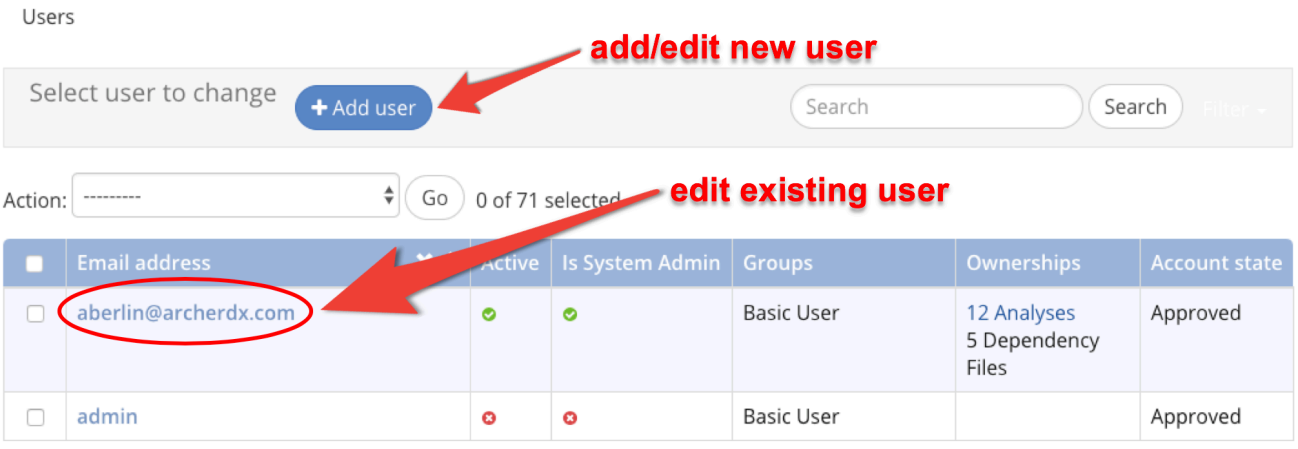


Multiple users can be selected by using the checkboxes in the left-most column, at which point they can be activated, deactivated, deleted, or approved/rejected all at once.

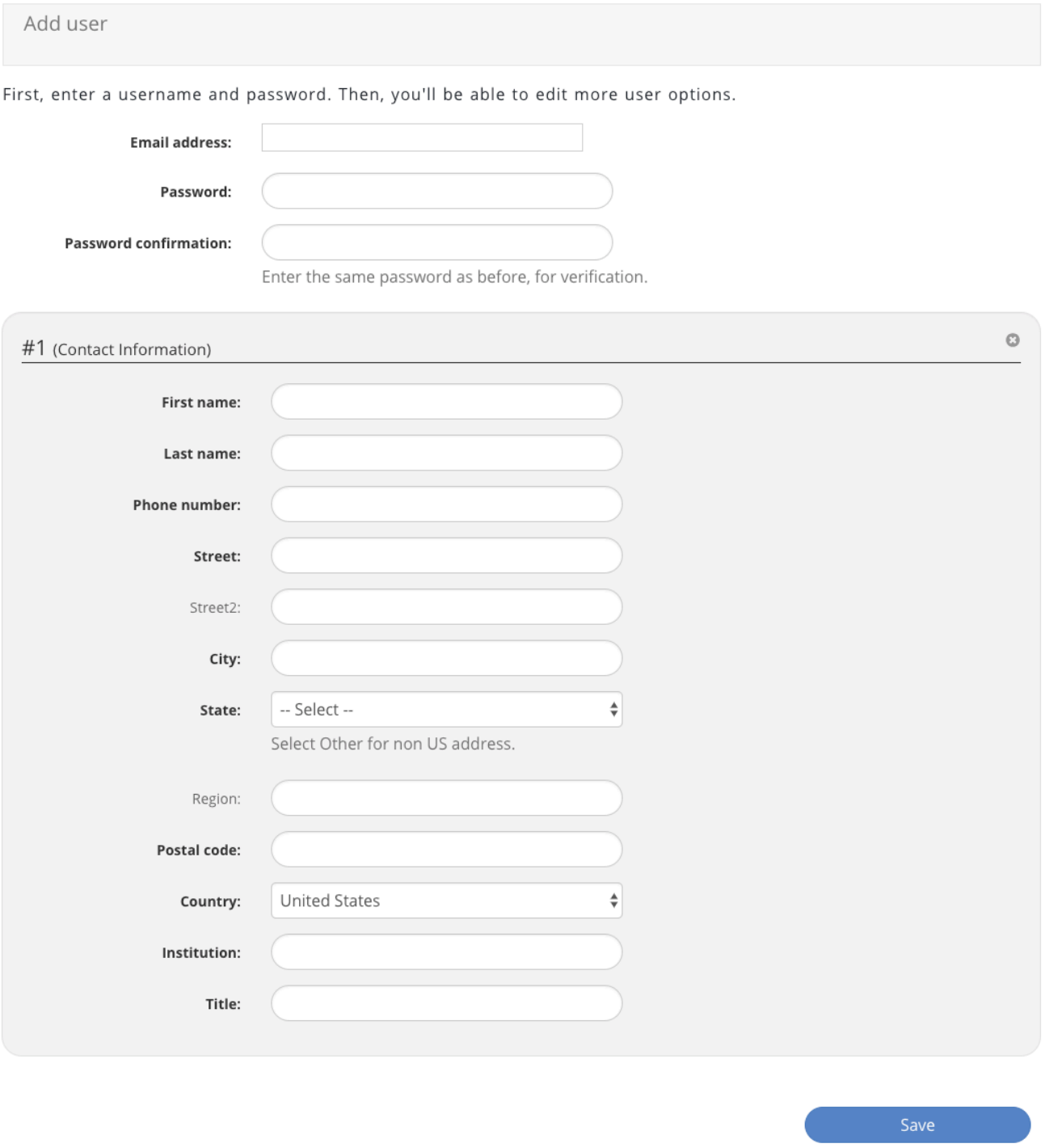
You can search for users by name by using the "**Search**" box.

Adding and/or Editing Users

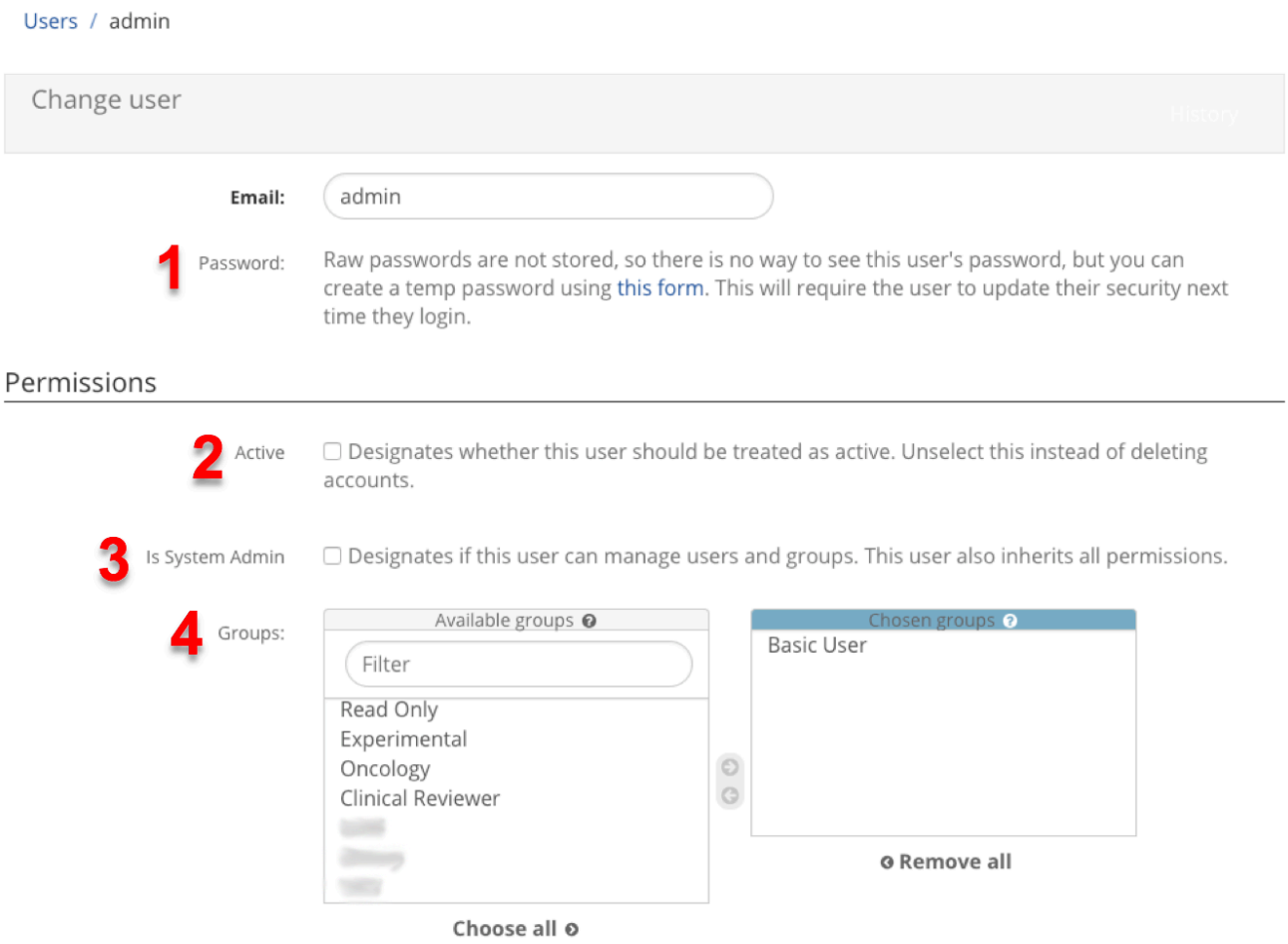
Click on the "**Add User**" button to go to the **Add User** page.

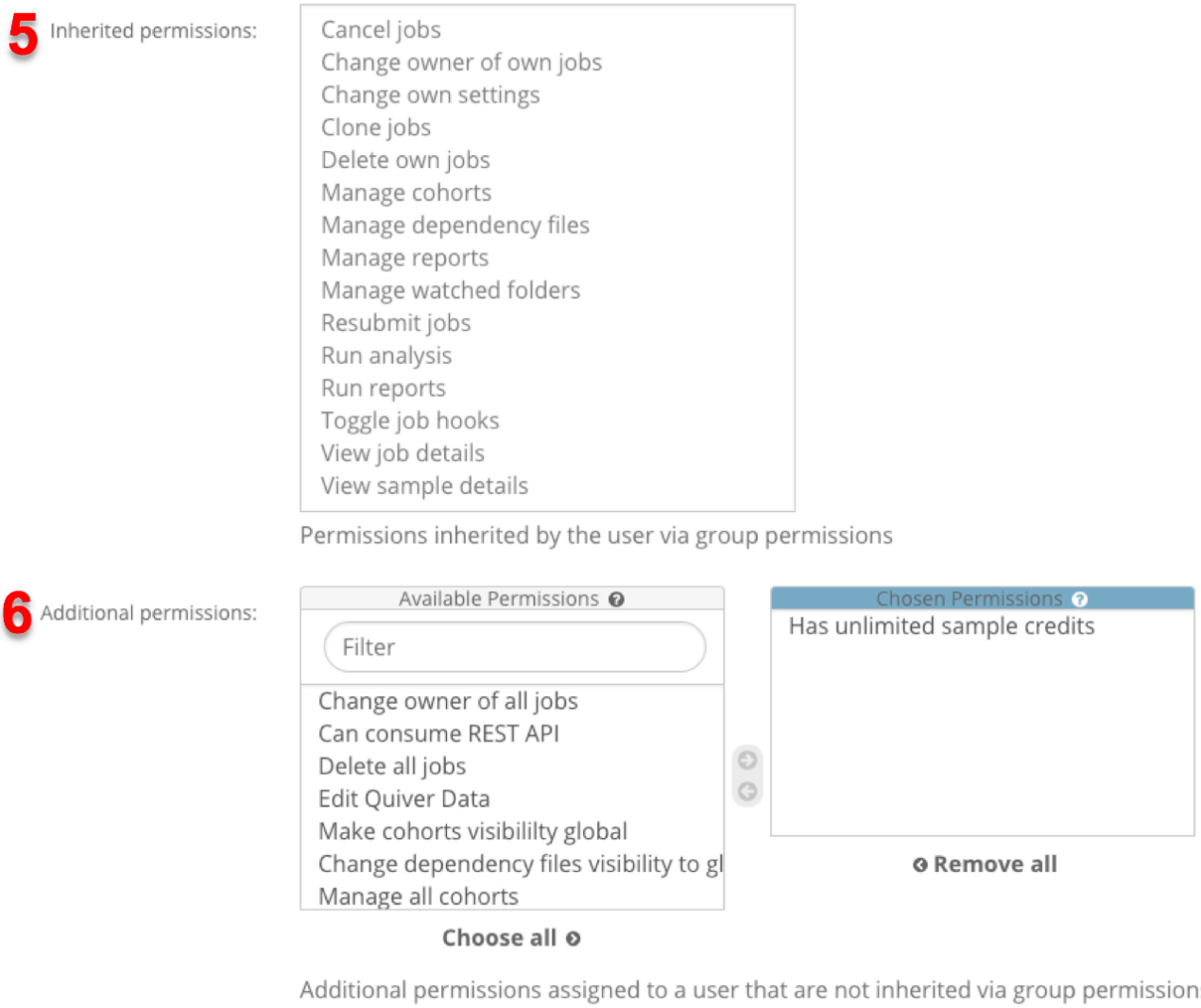


Provide the required fields and click the "**Save**" button:



You will then be taken to the **Edit User** page, where you can edit the new user. As shown in the screenshot above, the **Edit User** page can be reached by adding a new user or clicking on the username in the **Users** page:


Users / admin
Change user
History
Email:
admin
1
Password:
Raw passwords are not stored, so there is no way to see this user's password, but you can create a temp password using this form. This will require the user to update their security next time they login.
Permissions
2
Active
Designates whether this user should be treated as active. Unselect this instead of deleting accounts.
3
Is System Admin
Designates if this user can manage users and groups. This user also inherits all permissions.
4
Groups:
Available groups
Filter
Read Only
Experimental
Oncology
Clinical Reviewer
Choose all
Chosen groups
Basic User
Remove all
The groups this user belongs to. A user will get all permissions granted to each of their groups. Hold down "Control", or "Command" on a Mac, to select more than one.

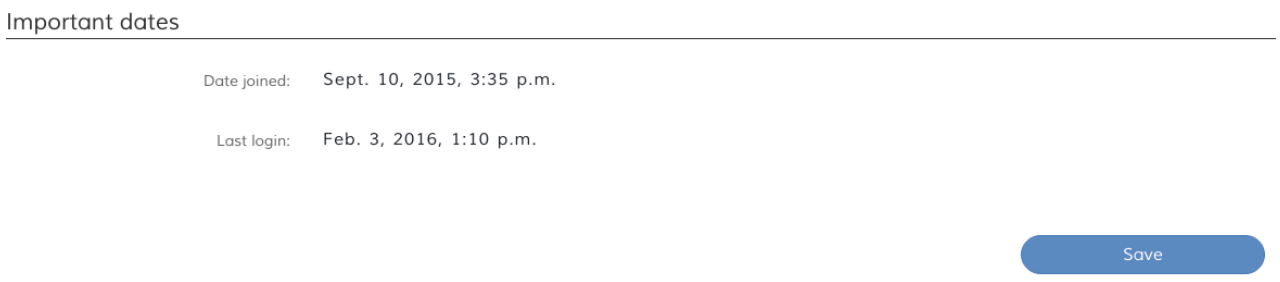


1. A new, temporary password can be assigned to the user.
2. Users can be deactivated by unchecking the "**Active**" checkbox.
3. Users can be granted the ability to manage users and groups by checking the "**Is System Admin**" checkbox.
4. Group membership is controlled by choosing Available Groups from the Group Selection List and moving them to the Chosen Groups list.

Detailed permissions can also be viewed and customized on the **Edit User** page, as shown below:

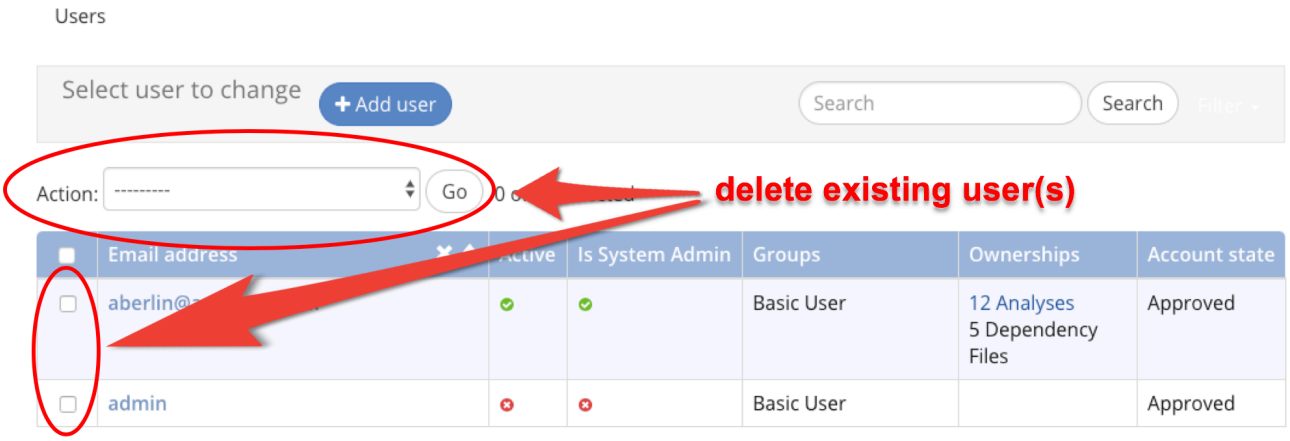
5. Permissions inherited by the chosen Groups are shown.
6. Additional permissions can be granted by choosing from the Available Permissions in the Permissions Selection List and moving them to the Chosen Permissions list.

When finished editing the user, you must click the "**Save**" button, located under the "Important Dates" section:



Deleting Users

Users can be deleted from the **Users** page by selecting the checkbox next to each username and then using the "Action" box:

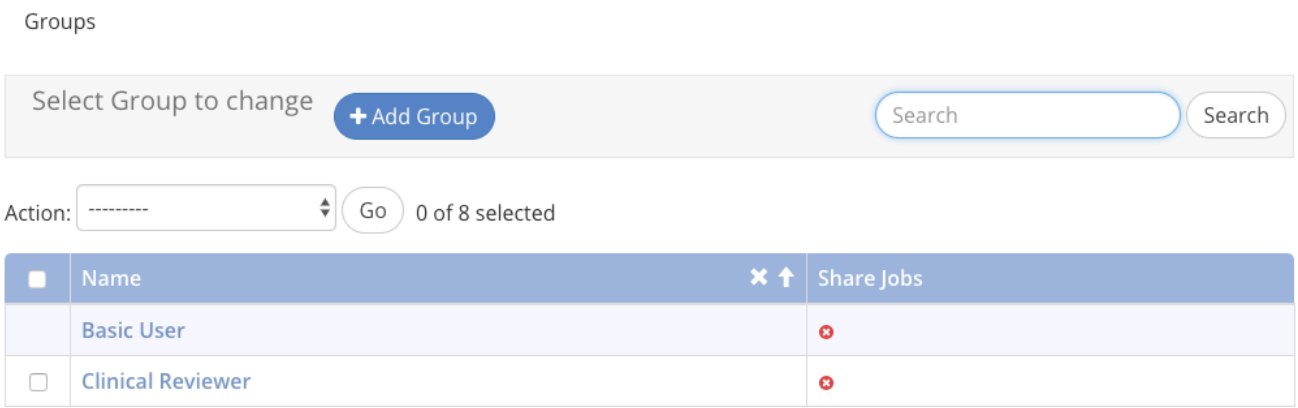


Groups

Groups serve two purposes: (1) conveniently manage groups of permissions, so that large numbers of individual permissions do not have to be manually assigned to individual users, and (2) allow users belonging to a given group to have visibility on the analyses performed by other group members, a functionality called Job Sharing.

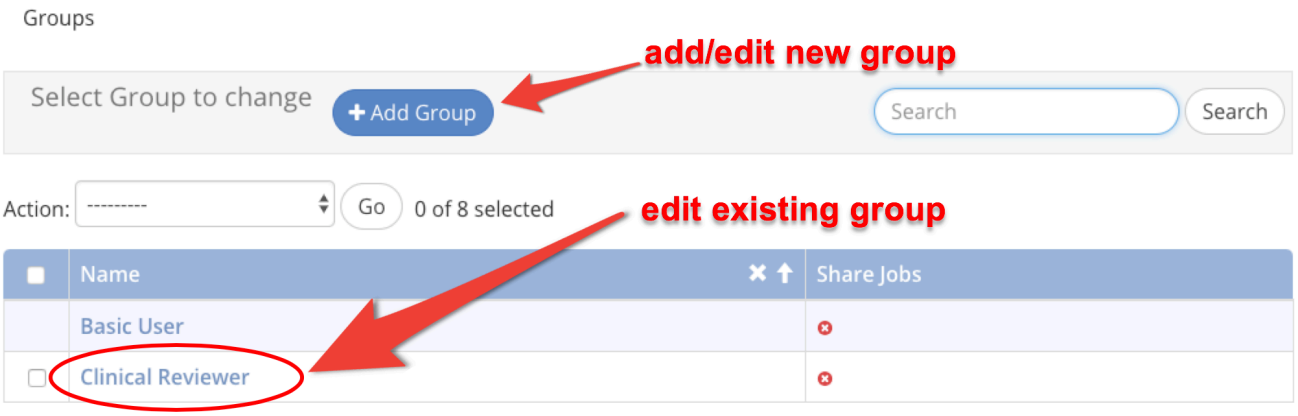
Select the "**Admin**">"**Groups**" menu item to manage all groups:

The **Groups** page shows a list of all the groups in the system. The columns indicate the name of the Group and whether Job Sharing is enabled for the group. You can find groups using the search box as well:

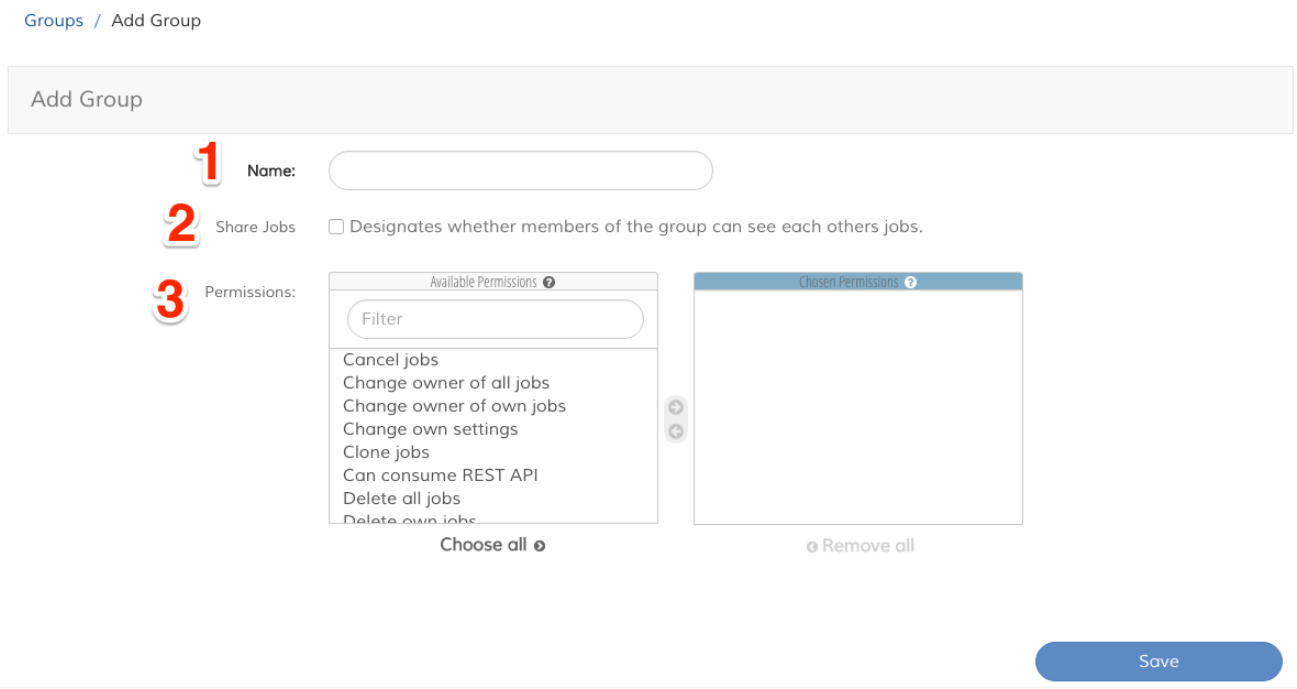


Adding and Editing Groups

Click on the "**Add Group**" button to go to the "**Add Group**" page:



You will then be taken to the "**Add Group**" page, where you can edit the group:

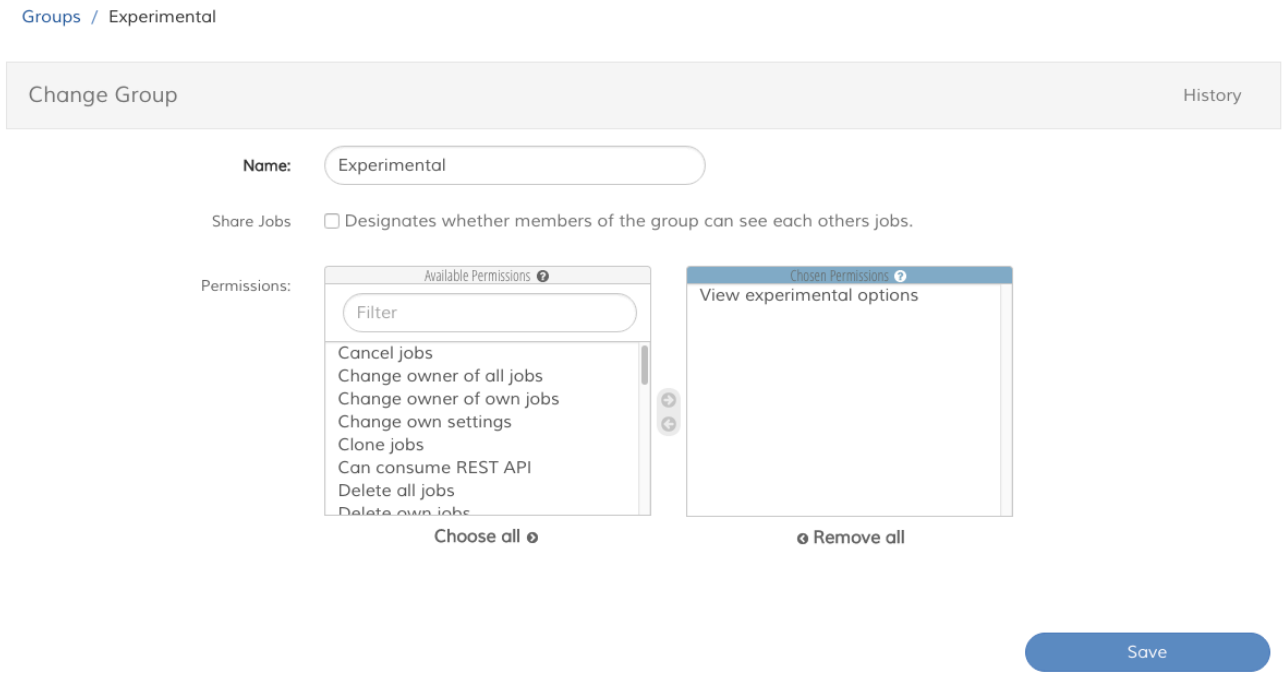


1. Enter a name for a the group
2. Select whether Job Sharing should be enabled for the group (i.e., all group members can see each others' jobs
3. Select specific additional permissions that should be enabled for the group

Provide the required fields and click the "**Save**" button.

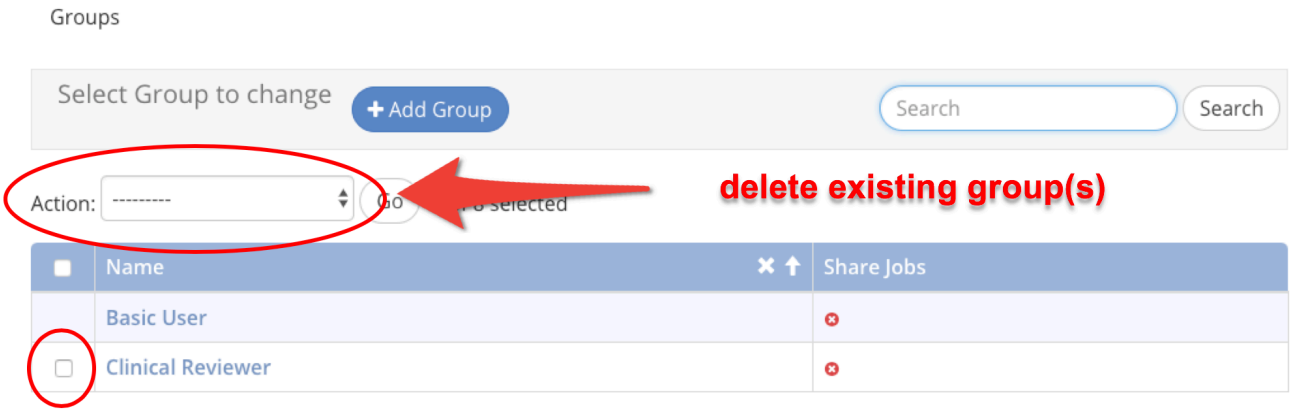
If Group Sharing is enabled, all members of the group will be able to see ***all*** jobs belonging to every member – including those intended for another group. Thus, best practice is to only assign users to multiple groups if they need to only view jobs, and not own any themselves. If a user is an admin, then they can run jobs and transfer ownership to other users.

As shown in the screenshot of the **Groups** page above, existing groups can be edited by clicking on the group name, which brings you to the **Change Group** page. The options on this page are identical to that of **Add Group** page:



Deleting Groups

Groups can be deleted from the **Groups** page by selecting the checkbox next to each username and then using the "Action" box:



Managing System (GTF and VCF) Files

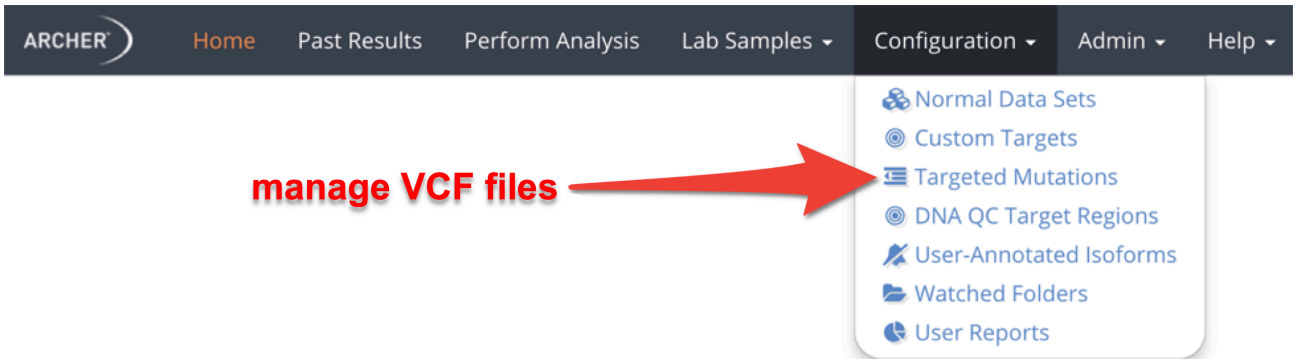
There are two types of files that are specified during the set up of an Analysis job:

- Assay Targets file, in Gene Transfer Format (GTF) – REQUIRED
- Targeted Mutations file, in Variant Call Format (VCF, v4.1) – OPTIONAL

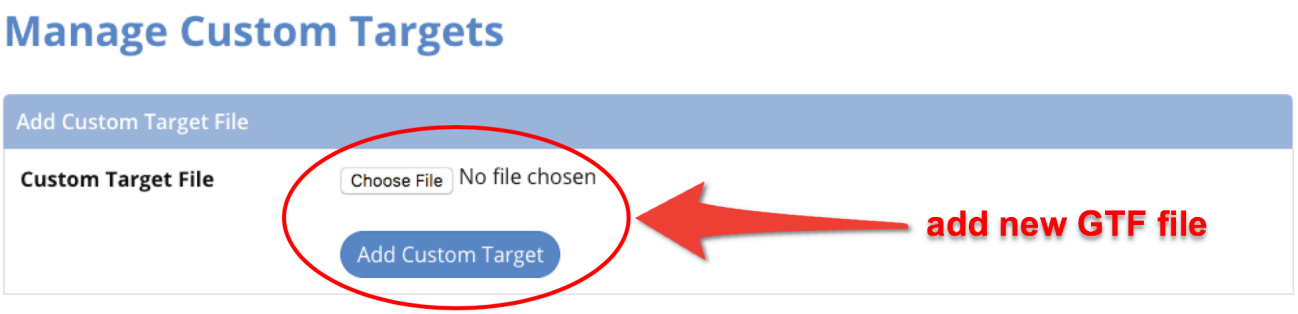
These files must first be uploaded and stored in Analysis before used in a job (several are already packaged with Archer Analysis for catalog panels). These files can be owned by the system, or by individual users (and in turn shared with all groups that the user belongs to). By default, new files are owned by the user that uploaded them, regardless of permission level. Admin users can manage GTF and VCF file ownership via the "**Configuration**" menu.

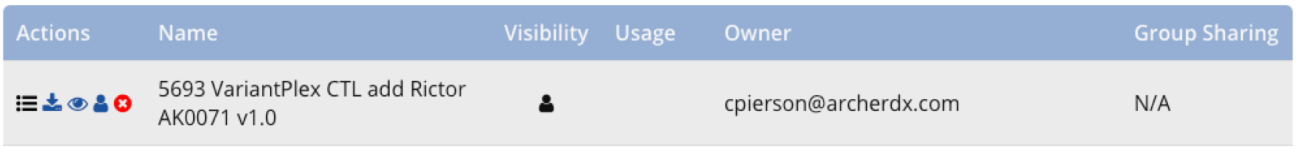
Managing and Adding GTF Files

Select the "**Configuration**">"**Custom Targets**" menu item to manage GTF files:



This will bring you to the **Manage Custom Targets** page, where you can add new GTF files, as well as view information about all existing GTF files in the system:





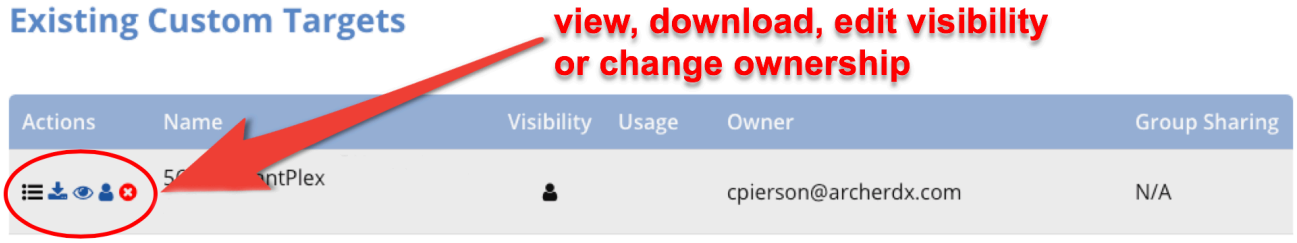
This page lists each GTF file name, visibility, usage, owner, and group sharing. '**Usage**' indicates how many jobs, if any, have used the file, as well as if the GTF file is part of a Watched Folders workflow (see Watched Folders for further details).

GTF File Actions (View, Download & Visibility)

Use options available under the 'Actions' column of the **Manage Custom Targets** page to (1) view the GTF file within the browser, (2) download a copy of the GTF file, or (3) edit the visibility:

Manage Custom Targets

Existing Custom Targets



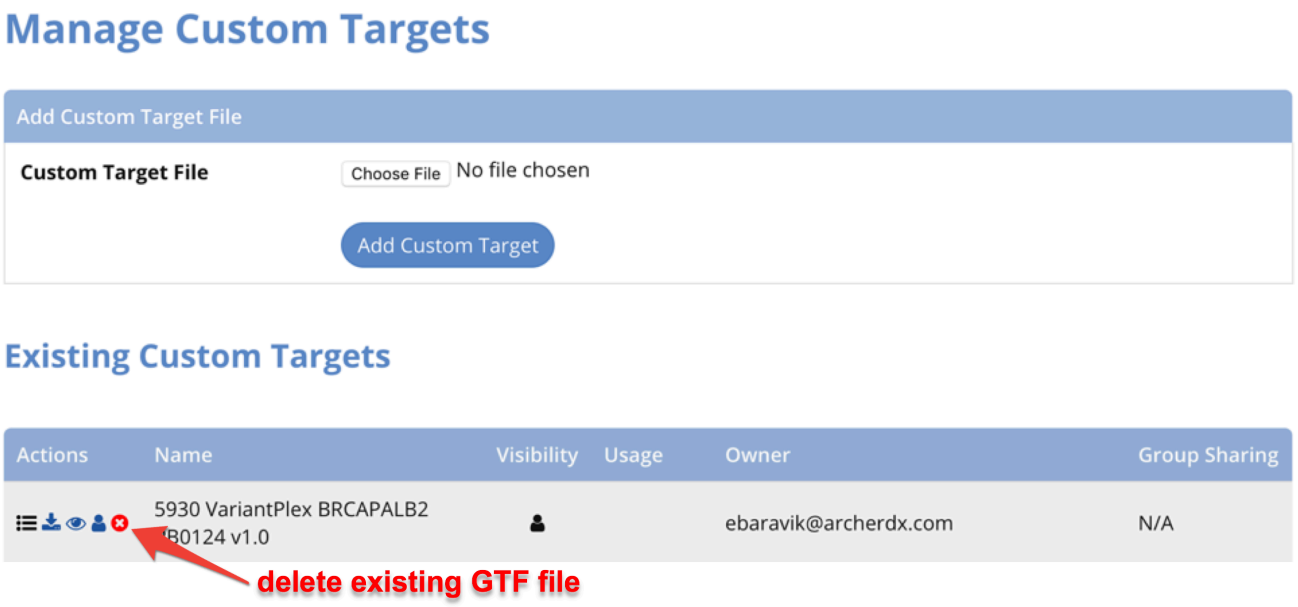
Selecting the 👁 will bring up a pop-up to edit GTF file visibility:

Select	Symbol	Name	Description
Current	👤	Owner	Only the owner will see the file.
Select	👥	Group	All users who belong to the same group as the owner which has sharing enabled will see the file.
Select	🌐	Global	All users on the system will be able to see the file.
Select	🌐	Global / Owned by System	All users on the system will be able to see the file and the file will not be associated to any particular user.
Select	👁‍🗨	Hidden	No one including the owner will be able to see the file when performing new analyses.

It is not possible to change ownership of GTF (and VCF) files to any user in the system, even as an admin. However, admin users can take ownership of a GTF file by first changing Visibility to ‘Global / Owned by System’ and then changing again to ‘Owner’.

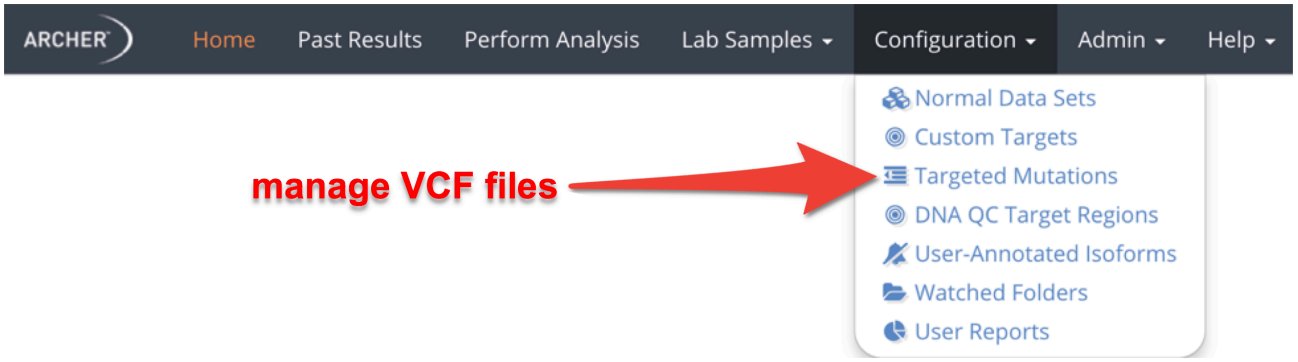
Deleting GTF Files

If a GTF file does not have any jobs associated with it, then it can be deleted on the **Manage Custom Targets** page:

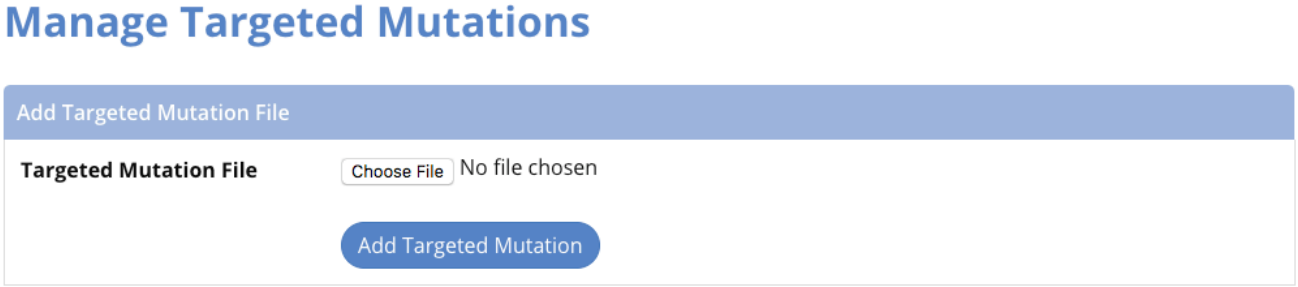


Managing VCF Files

Select the "**Configuration**">"**Targeted Mutations**" menu item to manage VCF files:



This will bring you to **Manage Targeted Mutations** page:



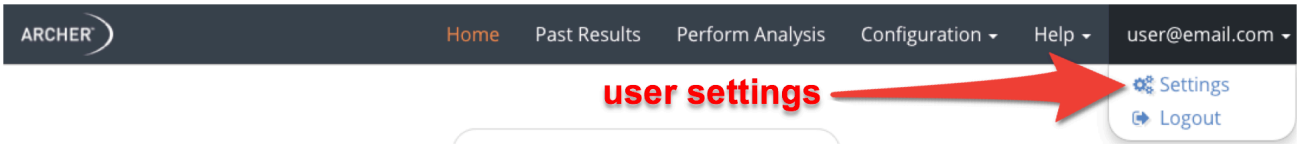
On the **Manage Targeted Mutations** page you can view, add, delete, download, edit visibility, change ownership and edit VCF files in a manner identical to GTF files (see "**Managing System (GTF and VCF) Files**").

User Settings (Applied Across All Jobs)

User settings apply to every future job started by a user. They cover parameters that effect both upstream and downstream points in the pipelines. The default values have been carefully selected, and should not be changed without extensive experience with the software, and/or consulting an Archer Bioinformatics Specialist (tech@archerdx.com).

Viewing and Changing Settings

User settings are managed on the **User Settings** page, which is accessed via the dropdown menu next to the username printed in the upper right corner of every page in the software:



User settings will only affect future jobs started by the current user. Changing settings will not affect prior jobs. If an admin changes its settings, it will not affect any other user in the system.

General Analysis Settings

The first section in the **User Settings** page contains parameters that identify core sample QC thresholds, or affect the very first steps of sample processing, i.e., will affect every pipeline available in the system:

MIN_AVERAGE_UNIQUE_DNA_AND_AMBIG_START_SITES_PER_GSP2

This is the key sample QC metric for the SNP/InDel pipeline that identifies the pass/fail criteria. It is defined as the minimum required average number of unique starts sites (note, **NOT** unique fragments) found across all GSP2's in the panel. The unique start sites are only those from reads that have been identified as DNA or ambiguous, and do not include RNA reads. (See *RNA vs. DNA vs. Ambiguous Reads* for further details on how reads are classified as RNA/DNA/ambiguous).

The default threshold is 50. This value is based on an in-house study aiming to achieve 90-95% sensitivity over a range of variants.

While the default value is a great starting point, it is highly recommended that a custom calibration be done at your own lab, in order to determine the optimum QC threshold that reflects the specific requirements of your lab.

MIN_AVERAGE_UNIQUE_RNA_START_SITES_PER_GSP2_CONTROLS

This is the key sample QC metric for the Fusion/Isoform pipeline that identifies the pass/fail criteria. It is defined as the minimum required average number of unique starts sites (note, **NOT** unique fragments) found across the control GSP2's in the panel. Control GSP2s in FusionPlex assays target genes that reliably express across any tissue type in the body. The unique start sites are only those from reads that have been identified as RNA reads. (See Key Assay Concepts for further explanation of unique start sites, control GSP2s, and how reads are classified as RNA/DNA/ambiguous).

The default threshold is 10. This value is based on an in-house study aiming to achieve 90-95% sensitivity over a range of fusions.

While the default value is a great starting point, it is highly recommended that a custom calibration be done at your own lab, in order to determine the optimum QC threshold that reflects the specific requirements of your lab.

ERROR_CORRECTION

This setting turns on/off the use of Molecular Barcode (MBC) based error correction. All reads that are associated with a single MBC are compared and condensed into a single consensus read, using the base call quality scores from primary analysis, and listed in the FASTQ.

Error correction is ENABLED by default. If false positives are of concern in your data, it is highly recommended that this setting remain ENABLED.

READ_DEPTH_NORMALIZATION

All samples in all jobs (regardless of pipeline(s) run) will be randomly subsampled to the number of reads specified in this setting. More reads per sample analysis requires longer processing time and more RAM. “0” represents NO normalization, and 3,500,000 is the default setting. Note, this metric setting is altered when running the ctDNA pipeline to 10,000,000.

BARCODE_LENGTH

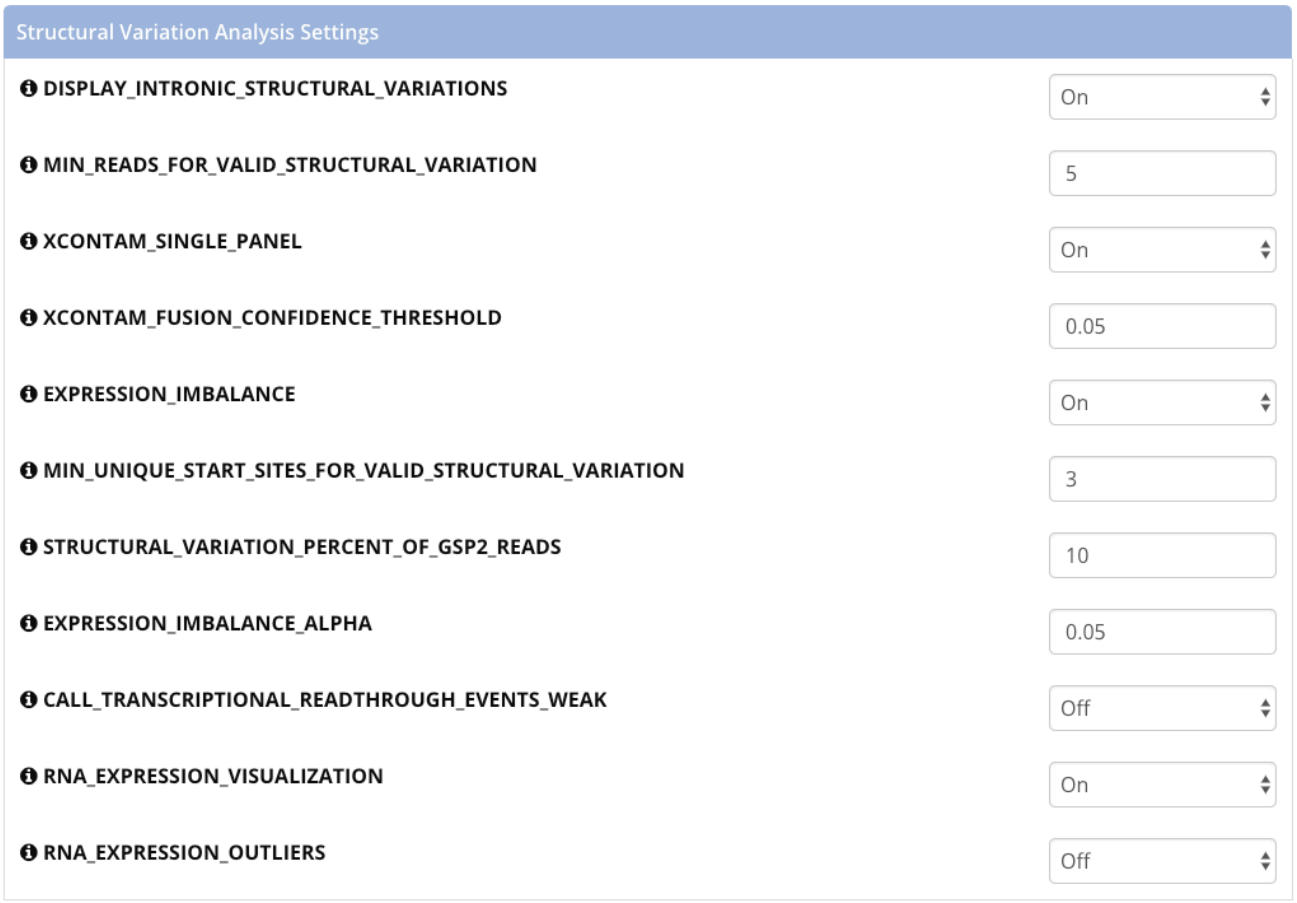
The number of nucleotides in molecular barcodes.

DEBUG

If DEBUG is turned on, all temp files will be kept within the job directory for debugging purposes. If turned off, temp files will be removed after the job is completed. DEBUG is OFF by default.

Structural Variation Analysis Settings

This section of the **User Settings** page contains parameters that affect only the Fusion/Isoform pipeline:



DISPLAY_INTRONIC_STRUCTURAL_VARIATIONS

This setting configures the pipeline to look for intronic structural variations.

MIN_READS_FOR_VALID_STRUCTURAL_VARIATION

This setting specifies the minimal number of breakpoint-spanning unique fragments (molecular bins) required to support a structural variation. Any fusion candidate that does not meet this threshold will not be evaluated further by the pipeline, and thus will not be displayed in the GUI or other file outputs.

XCONTAM_SINGLE_PANEL

This setting toggles ON/OFF the algorithm that can detect cross-contamination between samples run with the same panel (and run in the same job).

Cross-contamination is determined using a probability model, the basis of which is that duplicate fusions called across multiple samples (but within a single job) would also feature the same underlying read metrics. A scoring system incorporates the number of unique start sites, unique reads, raw reads, as well as average read depth (which is just the number of raw reads divided by the number of unique reads) for each fusion call. Each of these metrics is assigned a weight that is used to produce a final score (p-value). If that score surpasses a pre-determined threshold (empirically

determined by training the model on both in-silico and known contamination samples), then cross-contamination is flagged.

XCONTAM_FUSION_CONFIDENCE_THRESHOLD

This setting specifies the cross-contamination score (p-value) metric below which a Fusion Candidate will be flagged as likely contamination, and thus downgraded to the Weak Evidence bin.

EXPRESSION_IMBALANCE

This setting toggles ON/OFF the detection of expression imbalance across transcripts found from a given gene.

MIN_UNIQUE_START_SITES_FOR_VALID_STRUCTURAL_VARIATION

This setting specifies the minimal number of breakpoint-spanning fragments with unique start sites (a subset of unique fragments) required to support a structural variation. Any Fusion Candidate that does not meet this threshold will not be evaluated further by the pipeline, and thus will not be displayed in the GUI or other file outputs.

STRUCTURAL_VARIATION_PERCENT_OF_GSP2_READS

This setting specifies the minimal percentage of unique fragments from the GSP2's associated with a Fusion/Isoform Candidate that must be met in order to be classified as Strong Evidence. Otherwise, the Fusion/Isoform will be placed in the Weak Evidence bin.

EXPRESSION_IMBALANCE_ALPHA

This setting specifies the significance threshold for determining if a gene has expression imbalance due to the partial RNA expression of a fusion candidate. If below this threshold, the fusion candidate will feature an Expression Imbalance icon (⚖).

CALL_TRANSCRIPTIONAL_READTHROUGH_EVENTS_WEAK

This setting toggles the rule that, if a fusion candidate has been identified as a transcriptional read through event (as denoted by the ↻ icon), regardless of any other criteria, the fusion candidate will be placed in the Weak Evidence bin.

RNA_EXPRESSION_VISUALIZATION

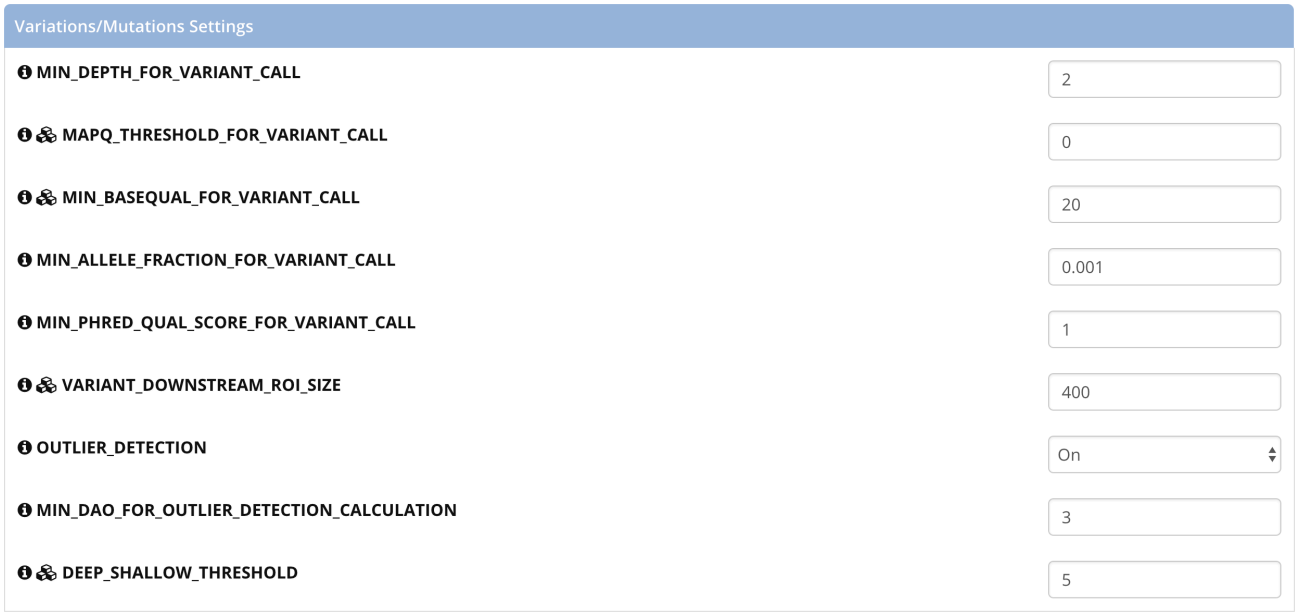
This setting turns ON/OFF the generation and presentation of a heat map visualization of the relative RNA expression for all samples in the run

RNA_EXPRESSION_OUTLIERS

This setting turns ON/OFF the detection of abnormally high RNA expression levels.

SNP/InDel Pipeline Settings

This section of the **User Settings** page contains parameters that affect only the SNP/InDel pipeline:



MIN_DEPTH_FOR_VARIANT_CALL

This setting specifies the minimum number of alternate observations (AO) supporting an allele in order to qualify as a potential variant (SNV or InDel) call. Any potential variants with AO below this threshold will not be reported by Analysis.

Note, this setting DOES NOT apply to Targeted Mutation calling (Vision).

MAPQ_THRESHOLD_FOR_VARIANT_CALL

This setting specifies the minimum MAPQ (Phred-scale) score of any AO supporting an allele in order to qualify as a potential variant (SNV or InDel) call. Any supporting reads with MAPQ below this threshold will not be used by Analysis to support a variant call.

It is recommended to leave this at the default value of 0, as other mechanisms of the SNP/InDel pipeline can be used to adequately screen for alignment and general call quality.

MIN_BASEQUAL_FOR_VARIANT_CALL

This setting specifies the minimum base call quality (Phred-scale) score of any AO supporting an allele in order to qualify as a potential variant (SNV or InDel) call. Any supporting AO with base quality below this threshold will not be used by Analysis to support a variant call.

MIN_ALLELE_FRACTION_FOR_VARIANT_CALL

This setting specifies the minimum allele fraction (AF = AO/DP) supporting an allele in order to qualify as a potential variant (SNV or InDel) call. Any potential variants with AF below this threshold will not be reported by Analysis.

Note, this setting DOES NOT apply to Targeted Mutation calling (Vision).

MIN_PHRED_QUAL_SCORE_FOR_VARIANT_CALL

This setting specifies the minimum variant call score (as reported by Freebayes or Lofreq) required to report the variant in the SNP/InDel pipeline. Any potential variants with a quality score below this threshold will not be reported by Analysis.

Note, this setting DOES NOT apply to Targeted Mutation calling (Vision).

VARIANT_DOWNSTREAM_ROI_SIZE

Specifies the number of bases downstream of the targeted GSP2s that are considered when looking for variants. Only applies when not already specified by the Target Region GTF file of the Archer assay. Targeted Mutations must be within the specified region of interest (ROI) to be reported.

OUTLIER_DETECTION

When enabled intra-job samples will be used to calculate the sequencing noise at a given position. The calculation will set the confidence in low AF variant calls.

MIN_DAO_FOR_OUTLIER_DETECTION_CALCULATION

The minimum number of Deep Alternate Observations on a variant required to perform Allele Fraction Outlier P-value calculations. Decreasing this parameter below 3 will increase analysis runtime.

DEEP_SHALLOW_THRESHOLD

The depth at which a molecular bin will be considered a 'deep' bin. Deep bins are considered higher confident reads as they have been assembled from multiple PCR duplicates.

CNV Pipeline Settings

This section of the **User Settings** page contains parameters that only affect the CNV pipeline in Analysis:

CNV_STRONG_AMPLIFICATION_THRESHOLD

This setting specifies the X-fold increase in Copy Number (CN), relative to the calculated baseline for the job, in order to be categorized as Strong Evidence Copy Gain.

CNV_STRONG_DELETION_THRESHOLD

This setting specifies the X-fold decrease in Copy Number (CN), relative to the calculated baseline for the job, in order to be categorized as Strong Evidence Copy Loss.

CNV_P_VALUE_THRESHOLD

In addition to the settings above, this setting specifies the confidence threshold (p-value) that a CN must be at or below, in order to be categorized as a Strong Evidence Copy Gain or Loss.

Miscellaneous System Settings

This section of the **User Settings** page contains parameters that either do not affect data analysis, or do not fit in the other page sections described above:

JOBS_PER_PAGE

This setting specifies the number of jobs that will display per page on the **Past Results** page.

LANDING_PAGE

This setting specifies the 'home' page of Analysis. The default is the **Running Jobs** page, as illustrated in The Home Page section above. However, this can also be changed to the **Perform Analysis** or **Past Results** pages, or the **Recent Samples** tab of the **Running Jobs** page.

CFDNA_READ_DEPTH_NORMALIZATION

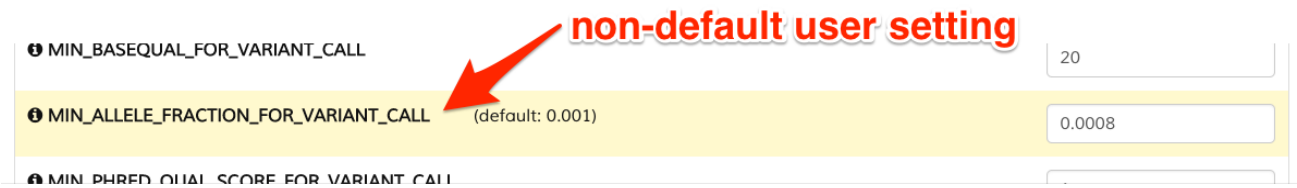
This setting overrides the READ_DEPTH_NORMALIZATION setting described above, for only jobs that run the 'cfDNA' option in the SNP/InDel pipeline.

Save New User Settings or Restore Defaults

After any changes are made on the **User Settings** page, remember to click the 'Update Settings' button at the bottom of the page. If it is desired to restore all original default settings in the system (i.e., as captured in the sections above), then instead select the 'Reset to Default Settings' button:

Identify Which Settings are Not Default

The **User Settings** page makes it very easy to determine which setting(s) have been changed from the default value(s), by highlighting those parameters in yellow, and including what the default value is:

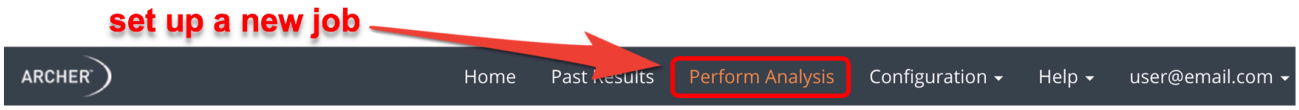


Job Setup and Management

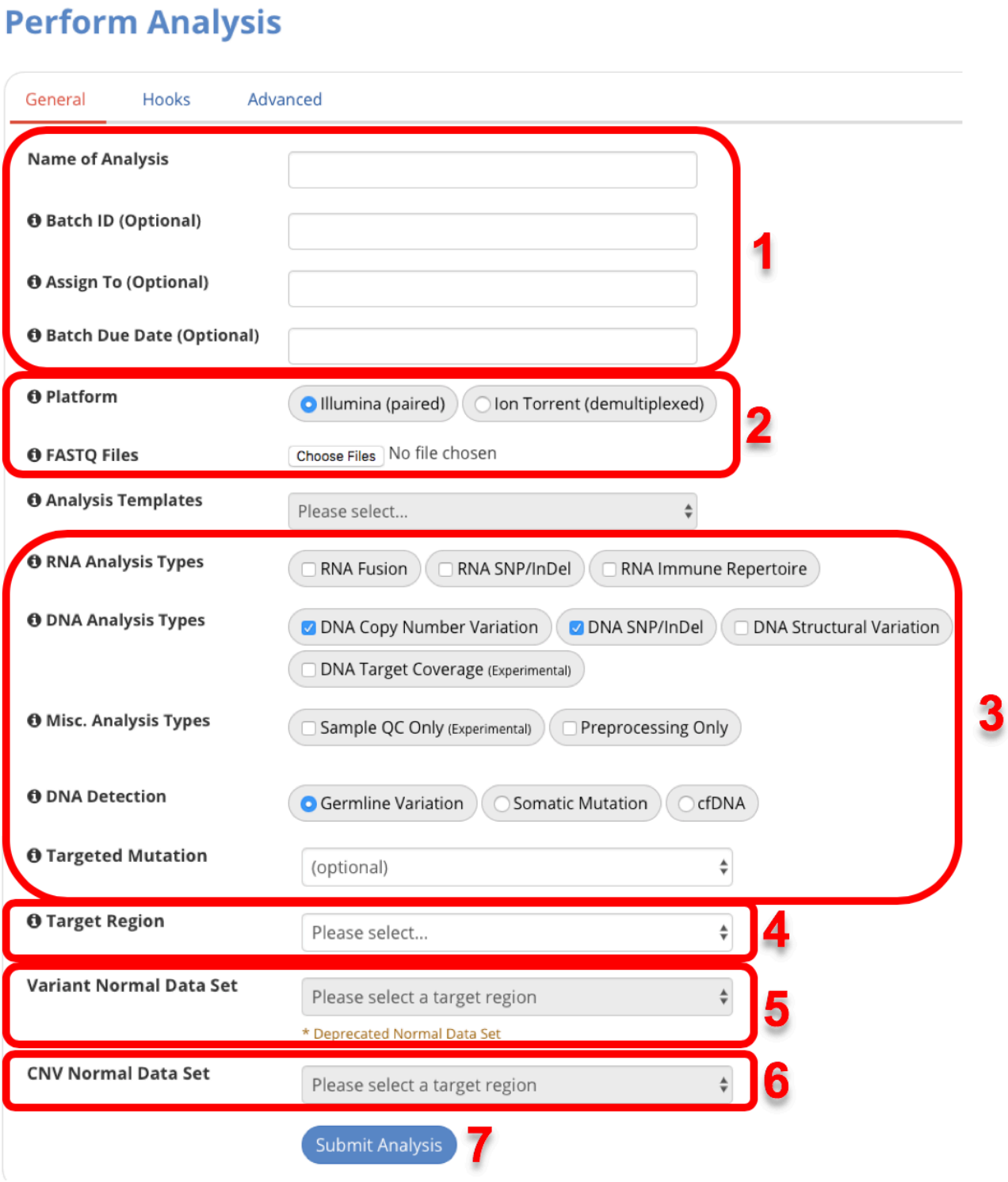
Jobs are fundamental to the Archer Analysis workflow. All samples must be processed as a part of a job, even if processed individually. All sample data is organized according to the job ID which is auto-assigned. Each job processes samples for a single Archer assay, but can include any number of samples, as well as run multiple pipelines in parallel (so long as they are compatible with the assay).

Starting a Job: Perform Analysis Page

Click on "**Perform Analysis"** in the Main Menu in order set up a new job:



This will bring you to the to the **Perform Analysis** page:


Perform Analysis
General
Hooks
Advanced
Name of Analysis
Batch ID (Optional)
Assign To (Optional)
Batch Due Date (Optional)
1
Platform
Illumina (paired)
Ion Torrent (demultiplexed)
FASTQ Files
Choose Files
No file chosen
2
Analysis Templates
Please select...
RNA Analysis Types
RNA Fusion
RNA SNP/InDel
RNA Immune Repertoire
DNA Analysis Types
DNA Copy Number Variation
DNA SNP/InDel
DNA Structural Variation
DNA Target Coverage (Experimental)
Misc. Analysis Types
Sample QC Only (Experimental)
Preprocessing Only
DNA Detection
Germline Variation
Somatic Mutation
cfDNA
Targeted Mutation
(optional)
3
Target Region
Please select...
4
Variant Normal Data Set
Please select a target region
* Deprecated Normal Data Set
5
CNV Normal Data Set
Please select a target region
6
Submit Analysis
7

1. *Enter key metadata for the job. At a minimum, a **Job Name** is required. To facilitate downstream analysis, the laboratory/sequencing **Batch ID** can be added (if not, the auto-assigned **Job ID** will be used here), and results review and reporting can **assigned to** any current user in the system. A **due date** for review/reporting can be specified as well. The Archer assay and pipelines used will be easily viewable in the results pages.*
2. *Select the platform and input files. If Illumina is chosen, then FASTQ files for all samples to be analyzed must be uploaded. Both R1 and R2 files must be selected for each sample. If Ion is chosen, then BAM files(s) for all samples must be uploaded.*

All FASTQ or BAM files to be analyzed in a job must be in the same folder, and selected at the same time. The files will be copied to and permanently saved in the system. Also, FASTQ files can be either uncompressed (with extension ".fastq" or ".fq") but it is ***recommended to use compression*** (using the GZIP algorithm, and have the extension ".gz"). ZIP compression is not supported.

3. *Select the Analysis Type(s), i.e. pipeline(s) used to analyze the samples.* Certain pipelines can be run together, and depending on which are chosen, additional setup steps may be required. See Choosing the pipeline(s) section below for further details.
4. *Enter the Target Region file.* As explained in Before Getting Started, this is the GTF file for the Archer panel used to prepare all samples to be processed in the job. All Archer catalog panels should already be listed. Custom panels require uploading of a dedicated GTF file, as illustrated in Managing and Adding GTF Files.

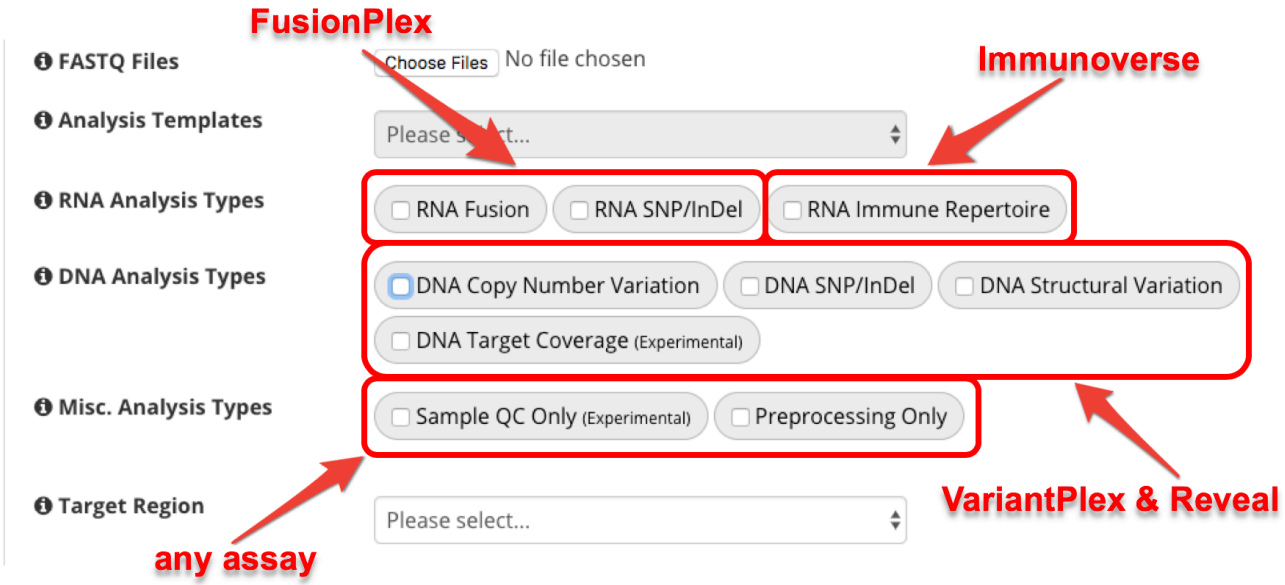
If the selected GTF file does not correspond to the Archer panel used to process the sample, there will be no automated flags from the system. However, analysis failure may be evident via the "On Target %" metric found in the Read Statistics tab of the **Sample Details** page. See Read Statistics for further details.

5. *Select Variant Normal Data Set* (optional). If a Normal Data Set was established prior to this analysis, it can be selected. Please refer to the "Normal Data Sets" section for more information.
6. *Select CNV Normal Data Set* (optional). If a Normal Data Set was established prior to this analysis, it can be selected. Please refer to the "Normal Data Sets" section for more information.
7. After all steps above are completed, click **"Submit Analysis"**, which will initiate upload of the sequencing files to the system. A progress bar, as shown below, will indicate the upload status. The speed of the upload depends on the network speed between the client and the Analysis server. If the client and the host are located on the same machine, the upload will progress very quickly, even for many GB of data. If the client is accessing a server that is on a network or on the Internet, the upload speed is determined by the network speed. The network upload speed can be verified with services such as http://speedtest.net.

After file upload is complete, the job will be entered into the queue, and processed as soon as system resources are available. At this point, the user can safely log out of the system.

Choosing the pipeline(s)

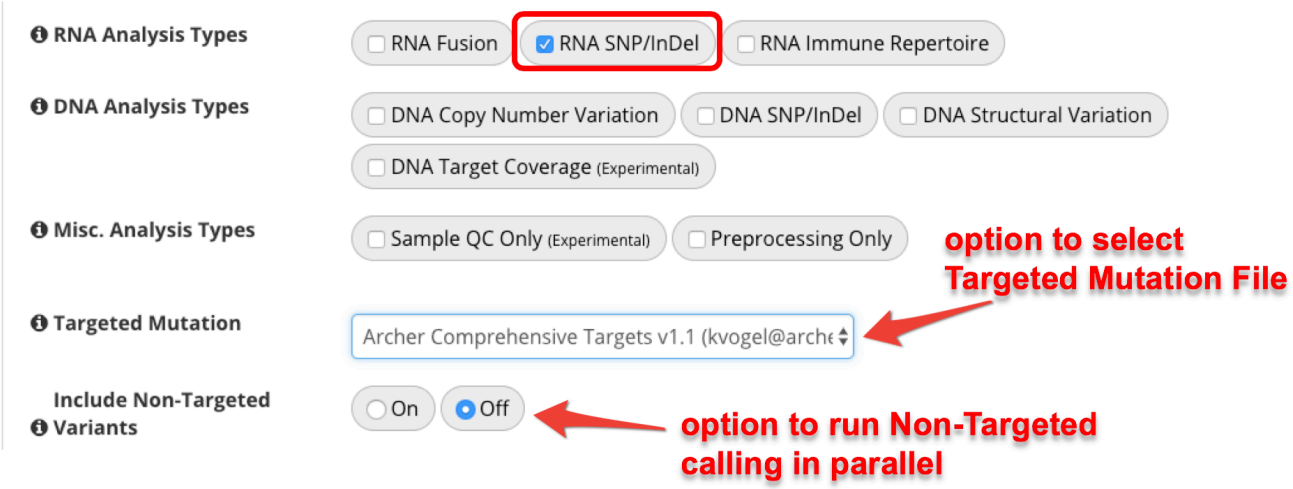
Depending on the type of Archer assay, a job may run multiple pipelines (a.k.a. Analysis Types). Only certain pipelines can be run together:



RNA Analysis Types

There are three pipelines for Archer RNA assays:

- **RNA Fusion** – select this to run the Fusion/Isoform pipeline for FusionPlex assays.
- **RNA SNP/InDel** – select this to run the Variants pipeline for FusionPlex assays. If selected, an additional field will display the option to select a Targeted Mutation File (TMF). Further, if a TMF is selected, another option will display to toggle Non-Targeted calling ON/OFF as well. Non-Targeted calling is what would run if no TMF were selected. See the Before Getting Started section of this manual for more information on TMFs:



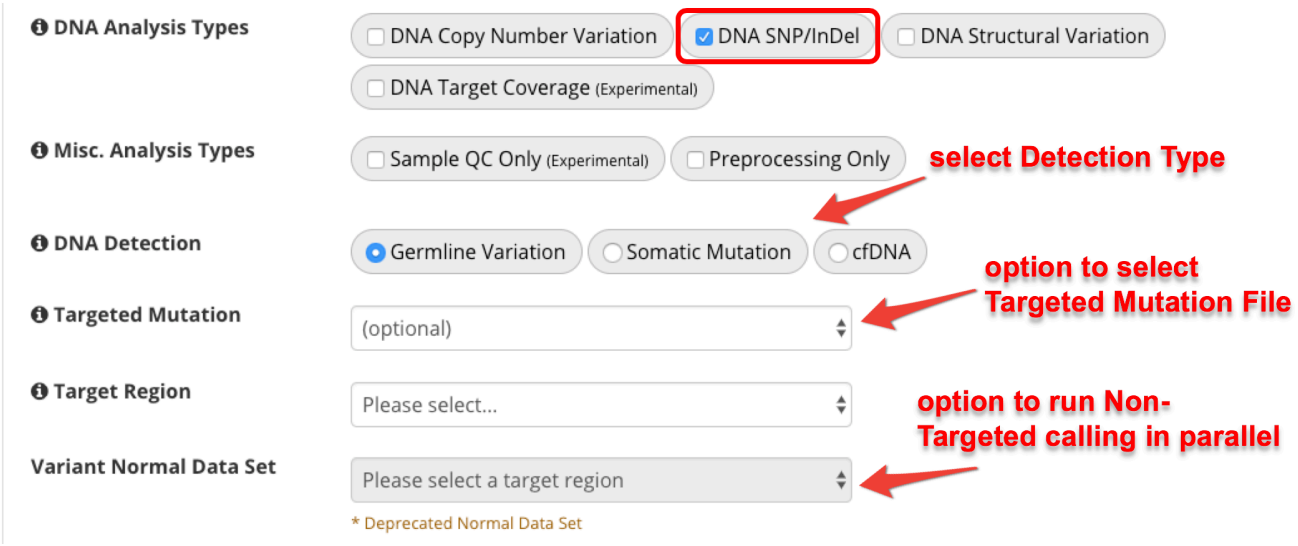
RNA Fusion and RNA SNP/InDel analysis can be run in the same job. Only exonic mutations will be detected with the RNA SNP/InDel pipeline. Both Germline and Somatic SNP/InDel detection are performed for RNA panels – see the DNA Analysis Types section below for more details on the specific SNP/InDel Detection Types.

- **RNA Immune Repertoire** – select this to run the immune profiling pipeline for Immunoverse assays. No other pipelines can be run in the same job (see Job Setup for Immune Repertoire for further details).

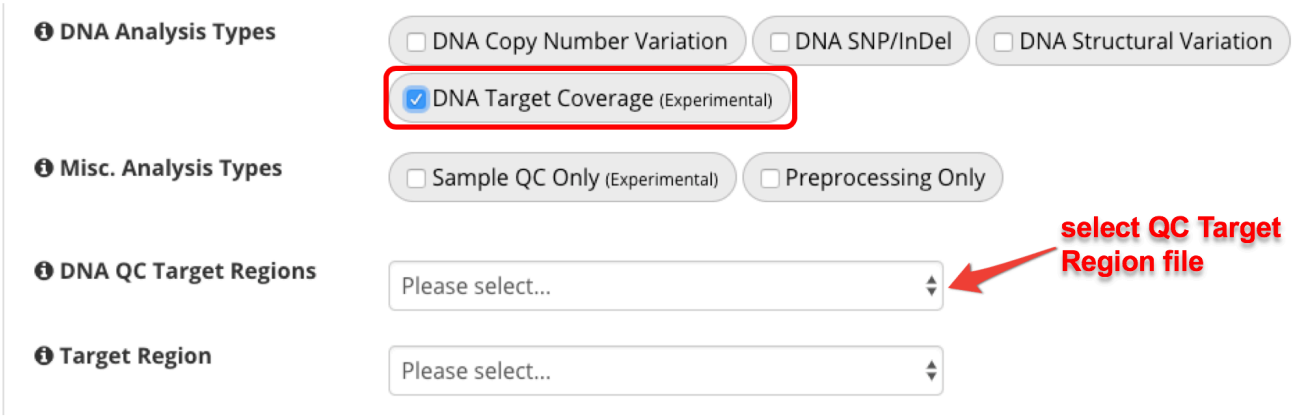
DNA Analysis Types

There are four pipelines for Archer DNA assays (VariantPlex and Reveal). All can be run in the same job:

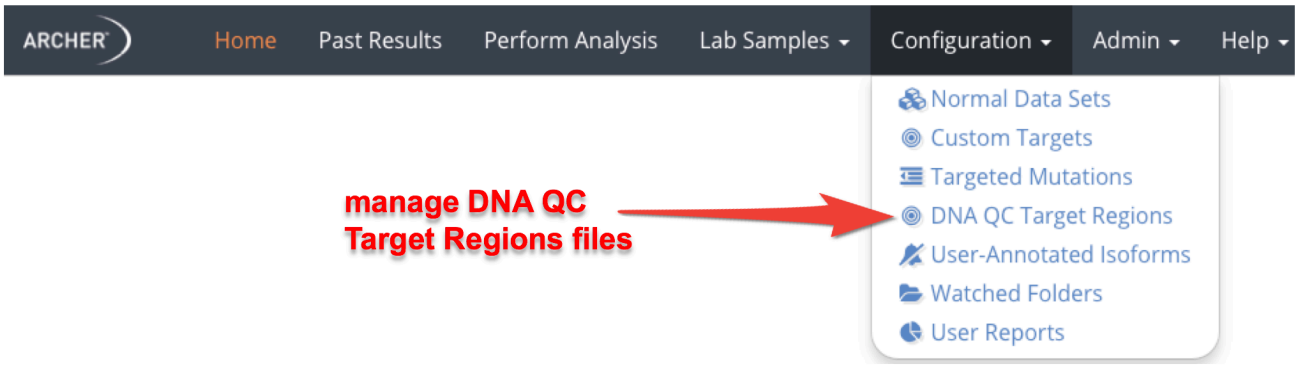
- **DNA Copy Number Variation** – select this to run the CNV pipeline. A further step is required to identify normalization parameters after "**Submit Analysis**" is clicked, as detailed in Selecting Normalization Parameters for CNV.
- **DNA SNP/InDel** – select this to run the Variants pipeline. If selected, an additional field will display to select the Detection Type (Somatic, Germline, or cfDNA – see DNA SNP/InDel Detection Types below for further details). Another field will also display for the option to select a Targeted Mutation File (TMF). Further, if a TMF is selected, another option will display to toggle Non-Targeted calling ON/OFF as well. Non-Targeted calling is what would run if no TMF were selected. See the *Before Getting Started* section of this manual for more information on TMFs:



- **DNA Structural Variations** – select this to run the Structural Variations pipeline, to detect large deletions, exon skipping events, internal tandem duplications (ITDs), and other structural rearrangements. This pipeline is analogous to the RNA Fusion/Isoforms pipeline, and utilizes de novo assembly.
- **DNA Target Coverage** – select this to run the experimental feature. See the *Experimental Features for Job Set Up* section for further details on this pipeline. If chosen, then an additional field will become visible for selecting a QC Target Regions file.

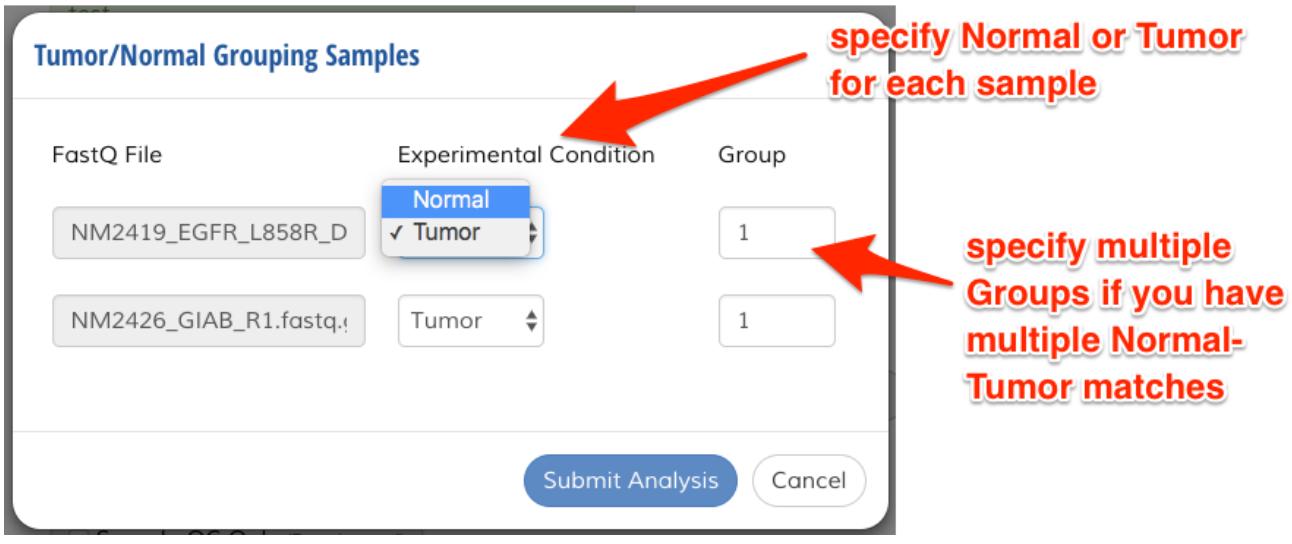


The DNA QC Target Regions file is in BED format and lists all the genomic regions for the pipeline to evaluate coverage. These files are managed in a manner identical to GTF files, via **"Main Menu ">"Configuration"** (see Managing System (GTF and VCF) Files section above for further details):



Selecting Normalization Parameters for CNV

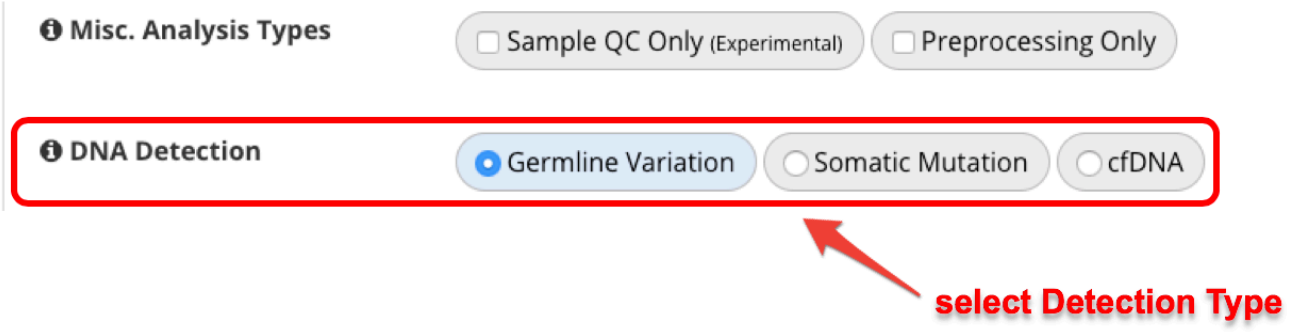
If the CNV pipeline is selected, a dialog box will appear (after **"Submit Analysis"** is clicked) that requires specification of how normalization in the algorithm is done for the job. This is done via selection of Normal versus Tumor samples in the job. CNV analyses will be more sensitive if a matched Normal sample is available for each Tumor sample (multiple Sample Groups can be identified within each job, if there are multiple Normal-Tumor matches):



Although recommended for maximum performance, it is not required to identify NORMAL samples in the job. Simply keep all samples labeled as TUMOR, and then all samples in each Group will be used collectively for normalization.

DNA SNP/InDel Detection Types

If the DNA SNP/InDel pipeline is selected, then a specific Detection Type must also be selected:



Germline Variation *– select this to detect only variants at relatively high allele fraction (AF). The germline variant detection algorithm will not reliably detect variants below AF of ~0.20. Also, the Germline filter set will be applied by default to the Variant Summary Grid, although this can then be changed at any time after the job is processed (see Filtering Results for further details).*

Somatic Variation *- select this to detect variants at any allele fraction (AF). The Somatic filter set will be applied by default to the Variant Summary Grid, although this can then be changed at any time after the job is processed (see Filtering Results for further details).*

cfDNA *– select this when running Archer Reveal panels; it is similar to Somatic Variation, except with advanced settings designed to improve sensitivity for very low allele fraction (AF) mutations in cell-free DNA from liquid biopsy assays. These settings are:* ***(1)*** *read depth normalization is set to 10 million (see Miscellaneous System Settings for customizing this parameter),* ***(2)*** *VARIANT_DOWNSTREAM_ROI_SIZE is set to 150bp,* ***(3)*** *Error Correction is turned ON, and* ***(4)*** *the cfDNA filter set is applied by default in the Variant Summary Grid (see Filtering Results for further details).*

For Reveal ctDNA panels: Make sure to run the cfDNA Detection Type in order to maximize sensitivity for very low AF mutations expected in these assays.

The Vision caller is always used for Targeted Mutations (i.e. if a TMF is selected) regardless of the Detection Type selected.

Target Region (GTF) File Annotations

Each primer (GSP2) listed in a panel's GTF file may be annotated to specify (1) which pipelines its reads can be applied to, and (2) the downstream Region of Interest (ROI) applicable to the SNP/InDel pipeline. Thus, any pipelines selected during Analysis setup may only apply to a subset of the reads

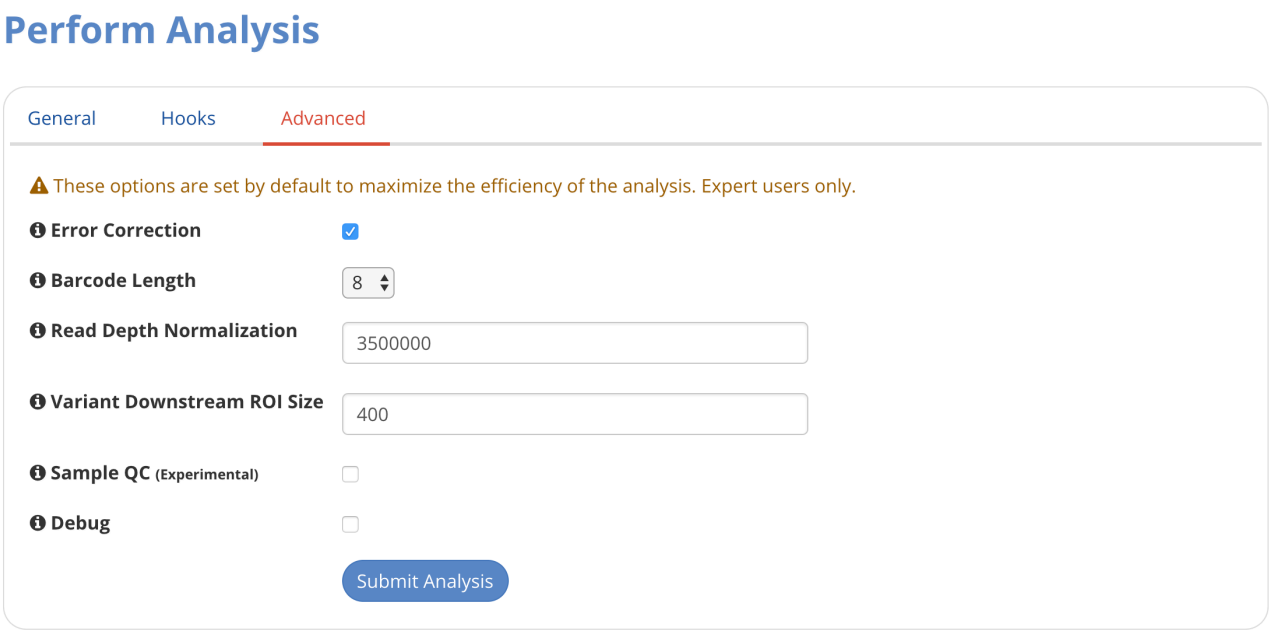
in the sample. These annotations are listed in the last column of the GTF, which can be viewed or downloaded by clicking the ☰ or ⬇ icons, respectively:

The last column of the GTF will list the annotations of interest. The **"function"** field lists the applicable pipelines for each primer. The **"target_ROI"** field specifies the applicable genomic coordinates.

Advanced Settings

Click **"Advanced"** on the **Perform Analysis** page to access advanced job settings:

This will bring up the "**Advanced Settings"** tab:



Several settings can be set here that will be applied for just the current job being set up:

- **Error Correction** – toggles Error Correction ON/OFF. See General Analysis Settings for more details on this feature.
- **Barcode Length** – The number of nucleotides in molecular barcodes.
- **Read Depth Normalization** – specifies the number of reads to subsample to. See General Analysis Settings for more details on this parameter.

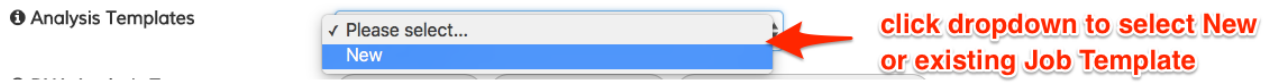
Variant Downstream ROI Size – specifies the ROI across all primers in the panel; will be overridden if specified per-primer in the GTF.

This setting turns ON/OFF the detection of abnormally high RNA expression levels.

- See *SNP/InDel Pipeline Settings* for more details on this parameter.
- **Sample QC** – toggles Sample QC user setting. See *Experimental Features for Job Setup* for more details on this function.

Analysis Templates

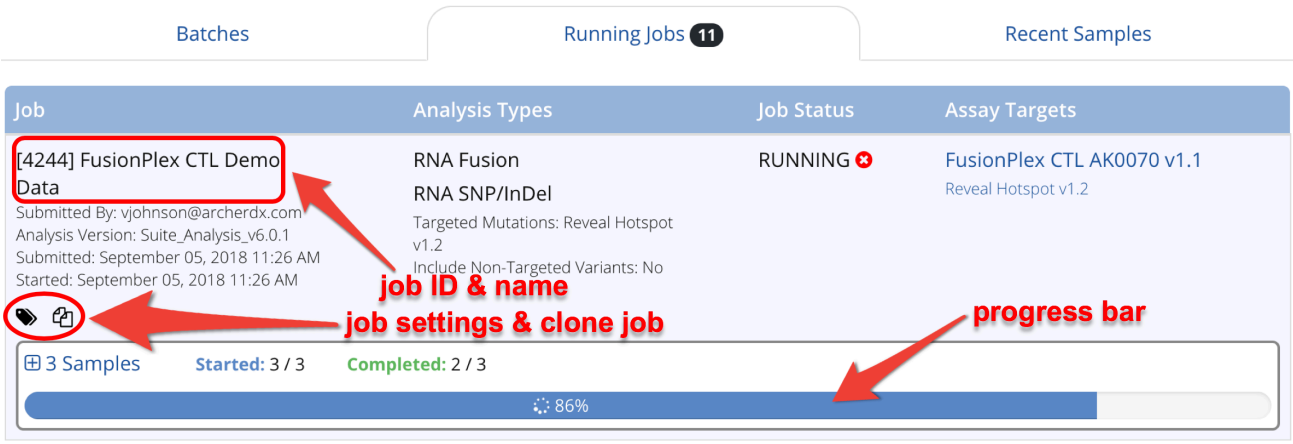
If you have a particular set of Analysis job settings that you would like to routinely run, you may save them as an Analysis Template by selecting **"New"** from the **"Analysis Templates"** pull-down list, and providing a name for it in the provided dialog box. Subsequently you can load it when setting up future analyses by selecting it from the "Analysis Templates" pull-down list.



In addition to all settings specified in the Perform Analysis page, the current User Settings will also be incorporated when defining a new Analysis Template. A useful trick is to create Templates as a shortcut for User Settings you wouldn't normally run across all jobs.

Running Jobs

Once a job as been submitted and sequencing files uploaded, it will be entered into the job queue, which can be viewed from the **Running Jobs** page (the default **Home** page – to change this, refer to Miscellaneous System Settings):



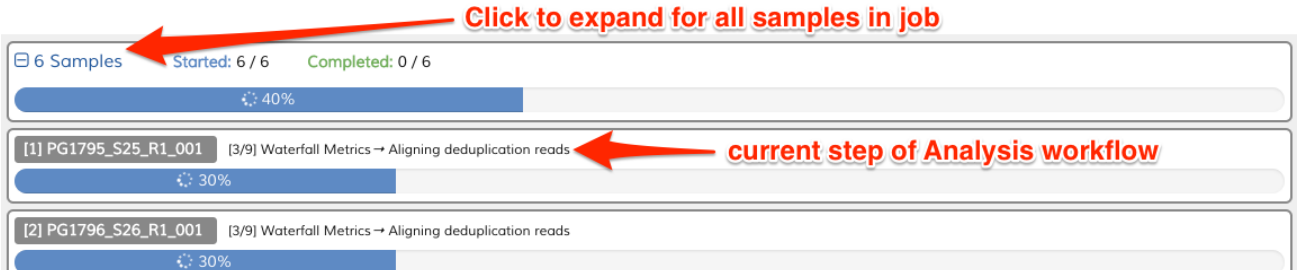
The **Running Jobs** page will display comprehensive information about every job currently in the queue, including Job Status, which can be: "NEW", "QUEUED", "RUNNING", "HALTED", "COMPLETED_ERROR" or "COMPLETED_OKAY".

From this view, the detailed settings for the job can be viewed by clicking on the 🏷 icon, and the job can be cloned by clicking on the 📋 icon. See Rerun or Clone a Job for further details on reanalyzing jobs.

DO NOT turn off the system or shutdown the machine hosting the Analysis server; this will disrupt currently running jobs, and most likely result in job failures.

Progress Bar and Status Updates

The progress bar indicates overall status of the job. This can be expanded to view progress bars for each sample in the job:



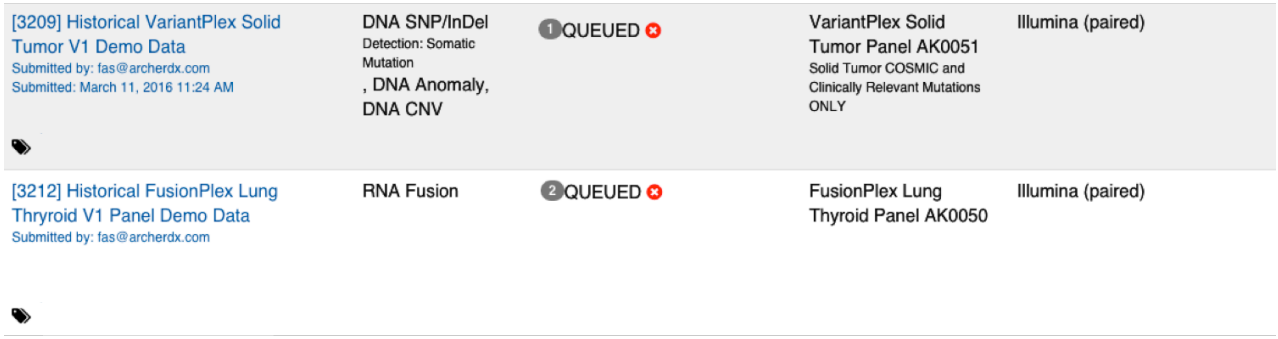
Below each sample's progress bar, the steps of the analysis workflow are indicated. In the example above, "[3/9]" indicates that step 3 of a total of 9 steps is being executed for sample [1].

Analyzing Multiple Samples & Jobs

The system can queue multiple samples simultaneously, either in a single job or split over multiple jobs. When a job has been started, another job can be initiated immediately without having to wait for

the previous job to finish. For local virtual machine (VM) deployments, the default situation is only one sample (and therefore job) will be running at once. Any new jobs that are started will be placed in a queue, as will the samples contained in that job.

When there are multiple jobs queued up, the location of the job in the queue will be indicated with a number as shown below. In this example, the job with the number "1" is the next in line to be executed when an available slot for the sample(s) within it becomes available.



Note that the VM can be configured to process >1 sample simultaneously if the VM is running on a host with sufficient memory. See the **Archer Analysis Virtual Machine Installation Guide** for further information.

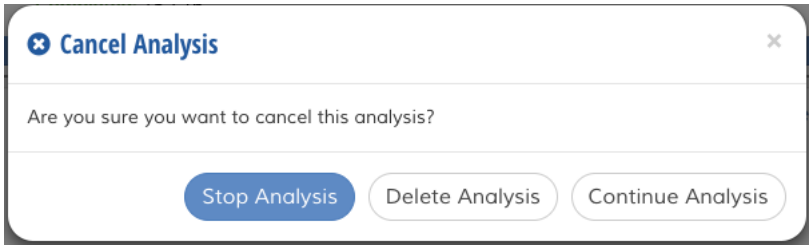
Stopping a Running or Queued Job

An analysis job that is queued or currently running can be stopped by clicking the icon:

There are two options for stopping a running job:

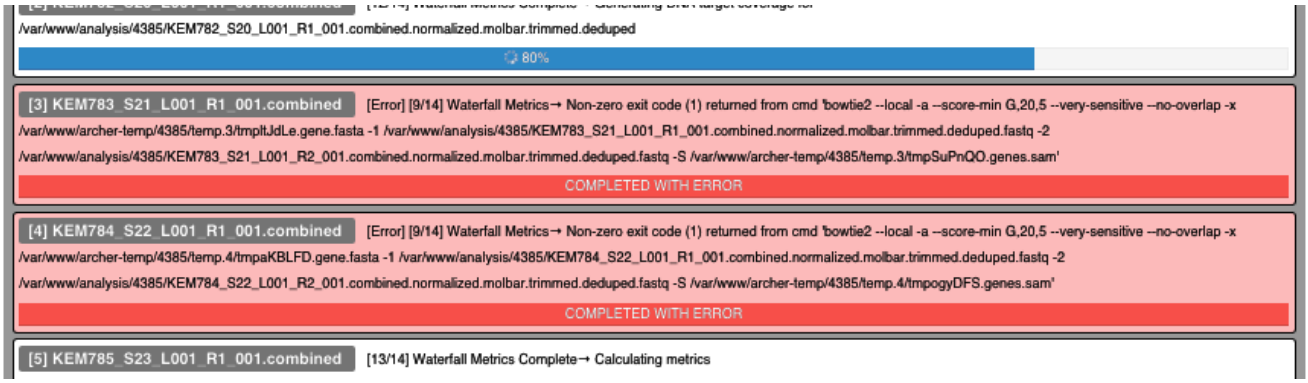
- **Stop Analysis** – will stop the job but leave the sequencing file(s) on the server and allow the samples to be rerun in the future.

- **Delete Analysis** – will stop the job and remove all the data and results from the server.



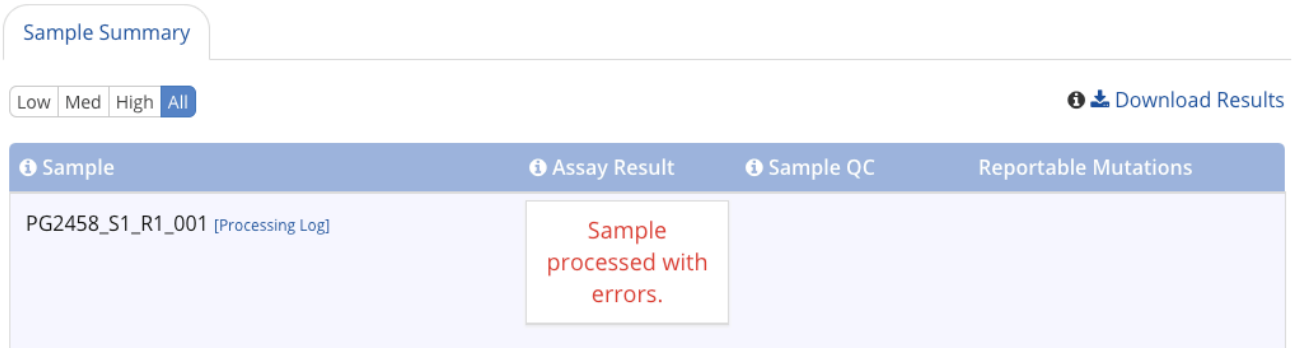
Jobs That Complete with Errors

It is possible that a job will finish with errors. If a job fails during a run, the sample(s) that fail will be indicated with a red bar.

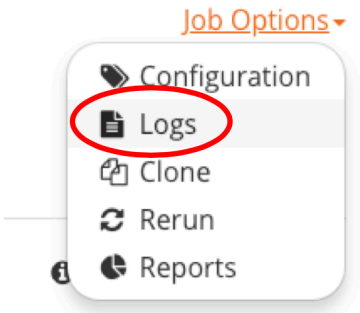


After all the samples have finished processing, the job will be marked as “COMPLETED_ERROR”. Consult the log files (icon) to determine the reason for the failed job and/or contact tech@archerdx.com for assistance.

In some cases, the error prevents the sample from being processed completely and this is indicated with the message “**Sample processed with errors**” in the **Job Details** page:



The number in the job queue ([3] and [4] in the example above) corresponds to the log file with the name “#.log.stderr.txt” (icon). Find that log file in the list of log files or click the “**[Processing Log]**” link in the **Job Details** page, or select “Processing Log” from the Options menu to download the log file to the local computer:



Job Archival States

Unarchived – Job is fully uncompressed locally. Job is fully viewable to users with the right permissions.

Archived – Job files that are non-essential to the frontend have been compressed. The number of days threshold to archive can be configured **on the frontend under admin –> System Settings** using the **DAYS_TO_JOB_ARCHIVAL** setting.

All jobs are archived after a default time frame of 30 days. Admin users have the ability to change this default setting to reflect the appropriate amount of time until a job should be archived. From "**Admin**">"**System Settings**", Admin users can change the "**DAYS_TO_JOB_ARCHIVAL**" file system setting. Set "**DAYS_TO_JOB_ARCHIVAL**" to 0 if you wish to disable the compression of your jobs.

Some features are disabled, such as:

File downloads + viewing

- Job Log Files
- CNV CSV Source File
- PDF Heat Map File
- VCF Summary Files

Visualization

- JBrowse Wildtype
- Isoform
- GSP2

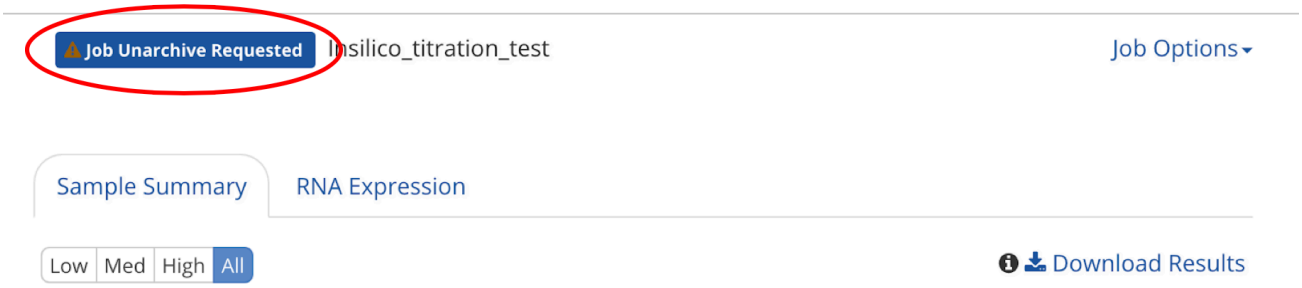
Archived in Glacier – Some Amazon Web Services (AWS) instances (Archer Analysis demo sites) have the option to move the Compressed Job file to AWS glacier from local storage. Files needed by the frontend persist locally, but job unarchival will take longer.

Once a job is in glacier, it can take up to 6 hours to restore the job, and there are data transfer fees, so glacier should only be used for long term infrequent data access

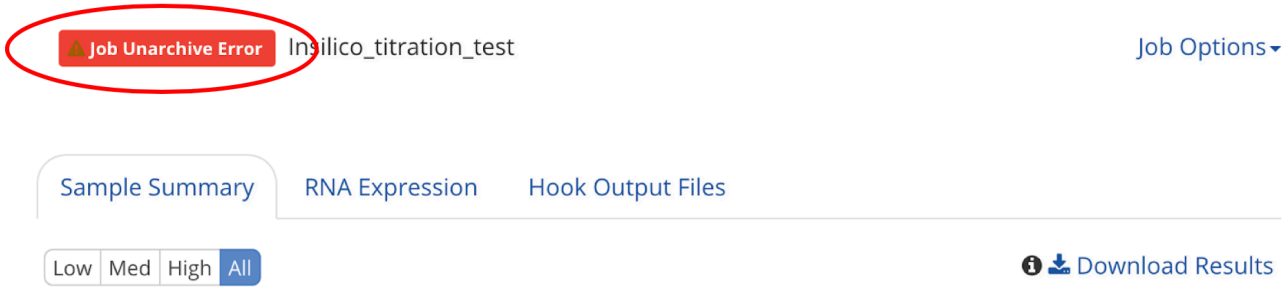
Unarchiving Jobs

Jobs that are archived can be manually unarchived from the "**Job Summary**" page by **Admin** users or users who have the **Managed Archived Jobs** permission.

Once a job unarchival is submitted, the unarchival button will turn blue and a box stating "**Job Unarchive Requested**" will appear.

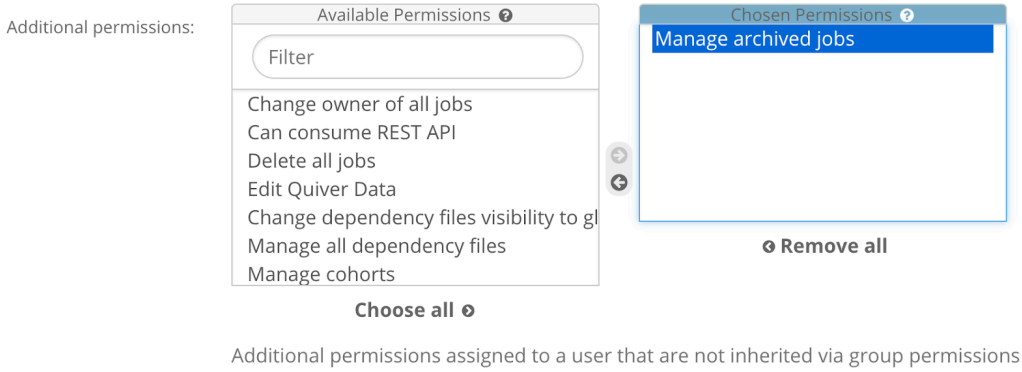


If there is an issue unarchiving a job, the unarchival button will turn red and a box stating "**Job Unarchive Error**" will appear.



Once the job unarchival is complete, the unarchive button will disappear. The job is fully uncompressed and users with the right permissions will have full viewability.

Admin users can grant basic users access to unarchiving jobs.



Enabling Glacier Archive – Glacier archive is disabled by default and to enable requires manual setup by Archer's development team.

Job Hooks

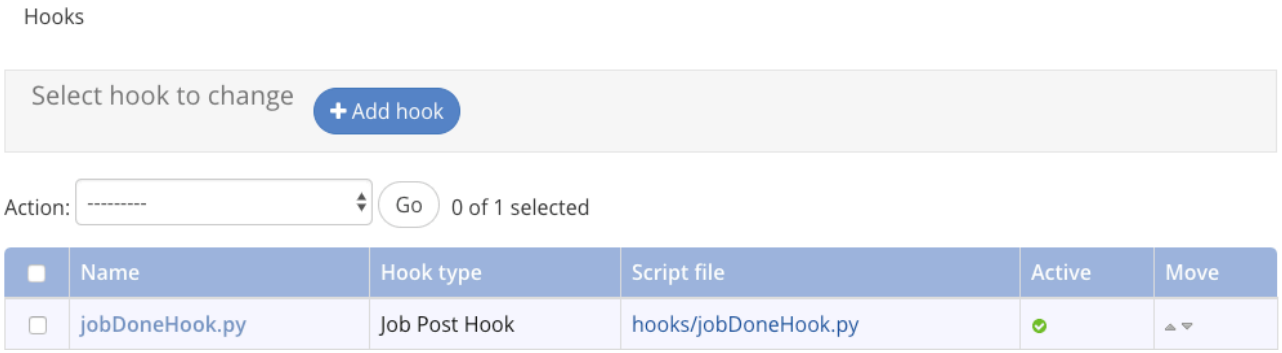
Job hooks allow users to execute custom scripts at specific points in the analysis pipeline. This mechanism allows the user to perform custom preprocessing steps, or custom tertiary analysis and reporting. The hook architecture provides the user with access to job identifying information and pipeline results through system level variables.

Please note that users must have Admin privileges to create job hooks; however, basic users may utilize job hooks once created.

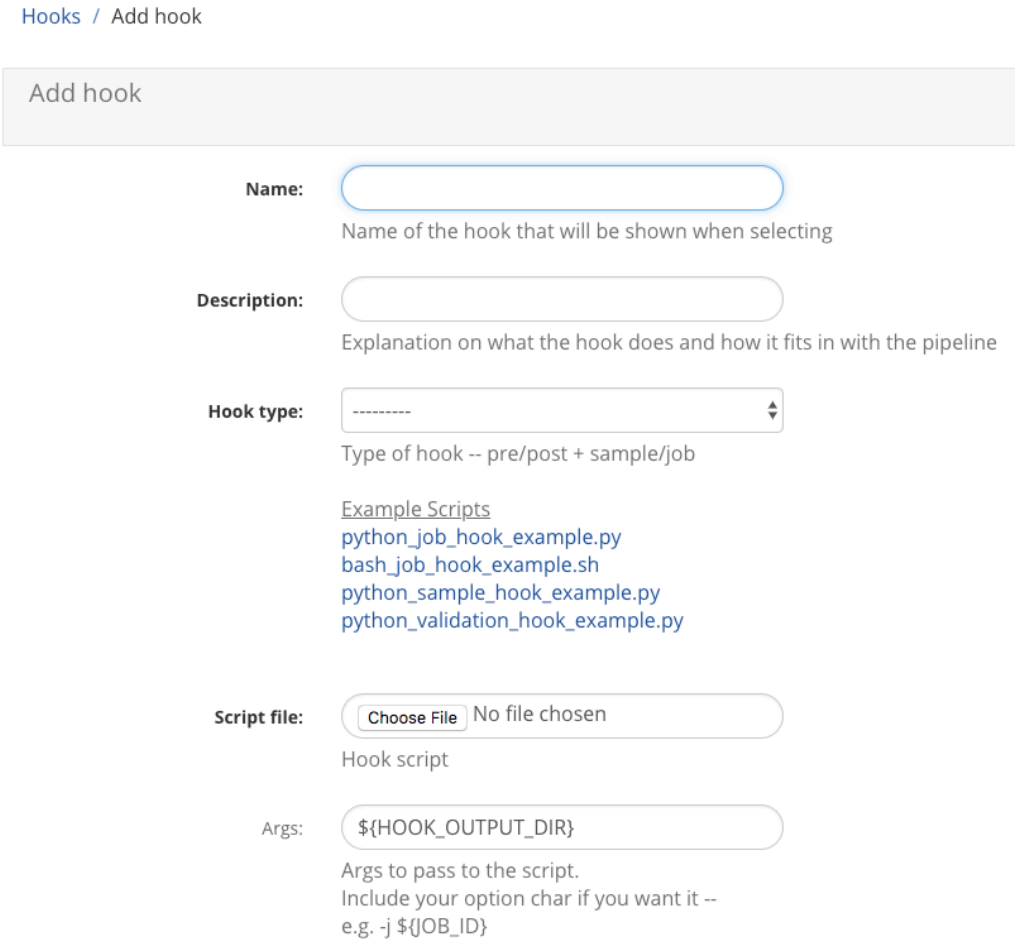
Select the "**Admin**"**>**"**Job Hooks**" menu item to add/manage all job hooks.

Adding a Job Hook

Click on the **"Add hook"** button to go to the "**Add hook"** page.



From the "**Add hook"** page, in order to create a new job hook, you will need to provide a name, description, hook type, script file and script arguments.



Hook Name: The hook name is important when differentiating between various hooks shown when setting up your analysis. The hook name must be unique to an analysis instance.

Hook Description: An explanation of what the hook does and how it fits with the analysis pipeline.

Hook Types: There are 5 hook types that define when the hook will be executed:

- **Job Pre Hook** – This hook type is implemented prior to a job running on the analysis pipeline.
- **Sample Pre Hook** – This hook type is implemented prior to each sample being run within a job in the analysis pipeline.
- **Job Post Hook** – This hook type is implemented after a job has completed the analysis pipeline.
- **Sample Post Hook** – This hook type is implemented after each sample has completed the analysis pipeline.
- **Validation Hook** – This hook type is used for software validation jobs being run on the analysis pipeline.

Hook script file: The script to be executed. This script can be implemented using Python version 2.7 or Bash (.py or .sh file extensions). Example hook scripts are located on the "**Add hook"** page and can be used as a base to develop a customized hook.

Script arguments: The variables allow information to be passed to the script. The only required system variable is **${HOOK_OUTPUT_DIR}**. All other system variables are optional.

- **Required:**
  - ${HOOK_OUTPUT_DIR} – This variable represents a location for output files to be displayed by the web UI**.**

- **Available Job Hook Variable Names:**
  - ${JOB_DIR} – This variable represents the job directory.
  - ${JOB_ID} – This variable represents the job ID.
  - ${SAMPLE_NAMES_COMMAS} – This variable represents sample names separated by commas.
  - ${SAMPLE_NAMES_SPACED} – This variable represents sample names separated by spaces.

- **Available Validation Hook Variable Names:**
  Optional arguments include the above job hook args, plus:
  - ${ORACLE_DATA_PATH} – This variable validates against expected results.

- **Available Sample Hook Variable Names:**
  - ${JOB_ID} - This variable represents the job ID.
  - ${SAMPLE_INPUT_DIR} – This variable represents the directory containing the input data.
  - ${SAMPLE_OUTPUT_DIR} – This variable represents the directory containing the output file for a sample.
  - ${PBS_ARRAY_ID} – This variable identifies the torque job scheduler ID.
  - ${SAMPLE_NAME} – This variable represents the sample name.
  - ${SAMPLE_ID} – This variable represents the sample ID.
  - ${LOG_FILE} – This variable represents a log file.
  - ${SEQUENCE_FILES_COMMAS} – This variable represents sample names, separated by commas.

- ${SEQUENCE_FILES_SPACED} – This variable represents sample names, separated by spaces.
- ${SAMPLE_VCF} – This variable represents the sample VCF file.
- ${FULL_RESULTS_FILE} – This variable represents the full results file.
- ${CNV_GENE_SUMMARY} – This variable represents the CNV gene summary file.
- ${CNV_PRIMER_SUMMARY} – This variable represents the CNV primer summary file.
- ${VARIANT_SUMMARY} – This variable represents the variant summary file.

To activate your newly created job hook, check the "**Hook enabled**" box next to the "**Active**" field on the "**Add hook"** page. Activated job hooks will execute for every job run on the system.

Once you have completed the required field, click the "**Save**" button.

Managing Existing Hooks

Editing Existing Hooks

Job hooks can be modified within the "**Job Hooks**" page by clicking the hook name you wish to edit. This will take you in the "**Change hook**" page where you can make modifications to the hook name, hook description, hook type, script file and hook arguments.

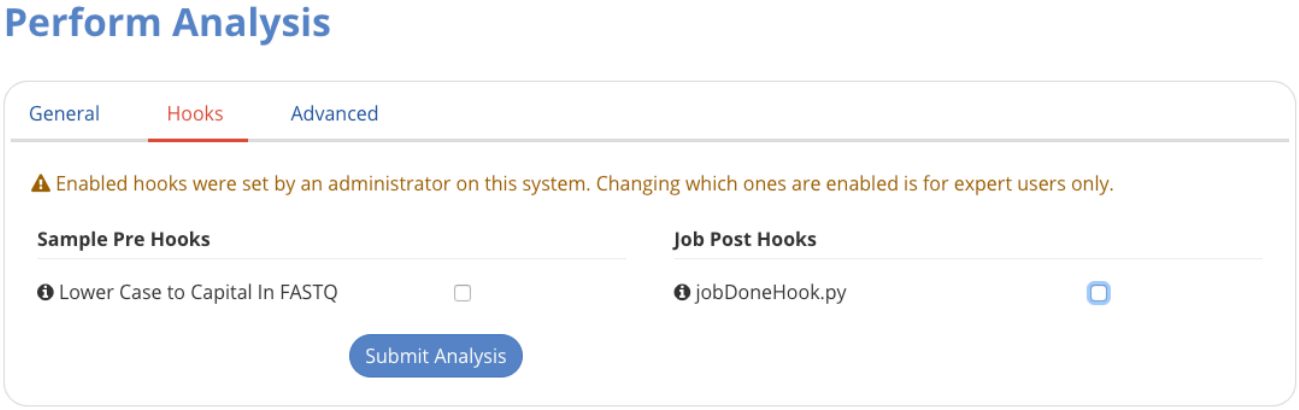
Disabling Existing Hooks

Job hooks can be disabled from within the "**Change hook**" page - from the "**Job Hooks**" page, click the job hook name you wish to disable. This will take you to the "**Change hook**" page. From there, scroll to the bottom of the page and make sure the "**Hook enabled**" box is unchecked.

Once unchecked, click the "**Save**" button.

Enabling hooks on a per job basis

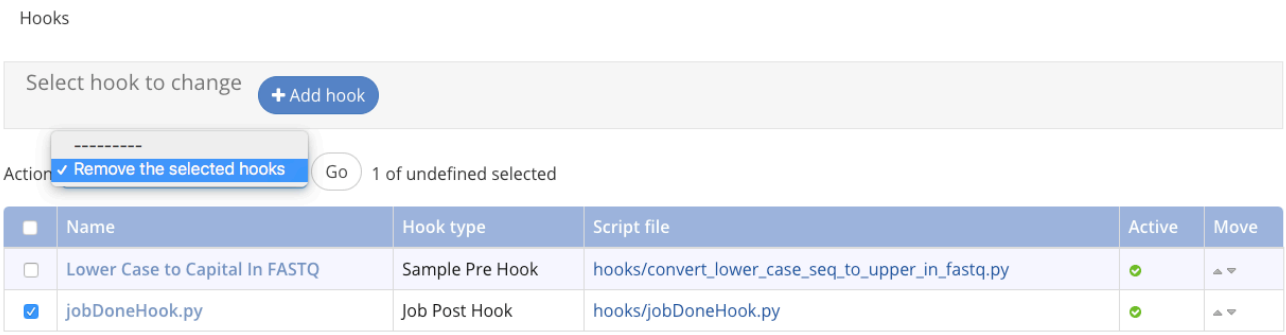
Job hooks can be enabled and disabled on a per job basis from the "**Perform Analysis**" page – when performing your analysis, you can enable/disable job hooks by selecting the "**Hooks**" tab on the "**Perform Analysis**" page and checking/unchecking all hooks that you wish to enable/disable.



Once all your other analysis settings are correct, click "**Submit Analysis**" to get your job started. Please note that this "**Hooks**" tab will not be visible to users unless one or more job hooks have been created. Please see "**Job Setup and Management**" section for more details.

Deleting Existing Hooks

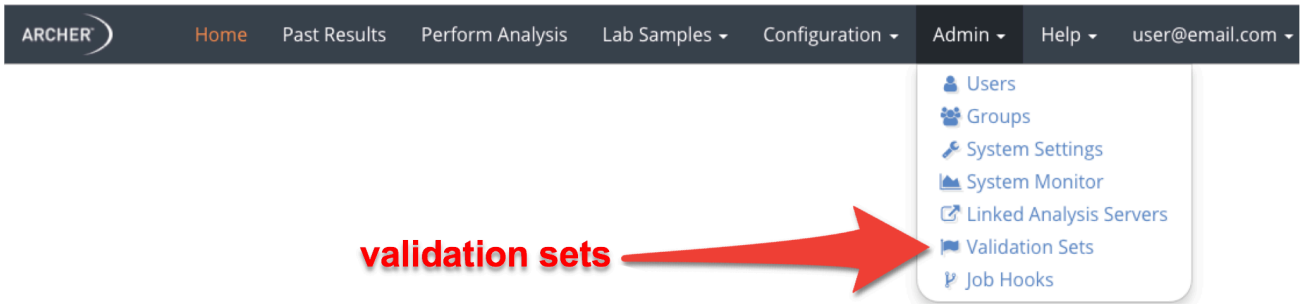
To delete a job hook, check the box to the left of the job hook name, then select "**Remove the selected hooks**" from the "**Action**" dropdown menu and click "**Go**".



Validation Sets

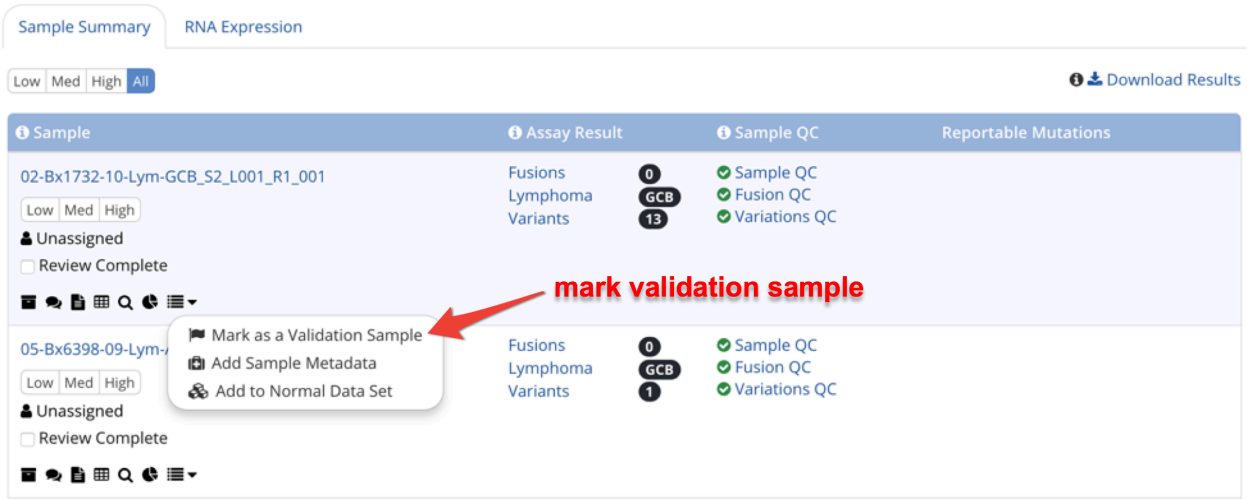
The use of validation sets is important to users that wish to validate the version of Archer Analysis that they are running. Users are able to select any of their completed samples to produce a validation set that executes and automatically checks to confirm the software is working properly. This feature not only notifies the user that a validation set passed/failed, but a detailed PDF report can also be downloaded. Note, only Admin users have the ability to run validation sets.

To access Validation Sets, select the "**Admin**"**>**"**Validation Sets**" menu to add/manage Validation Sets.



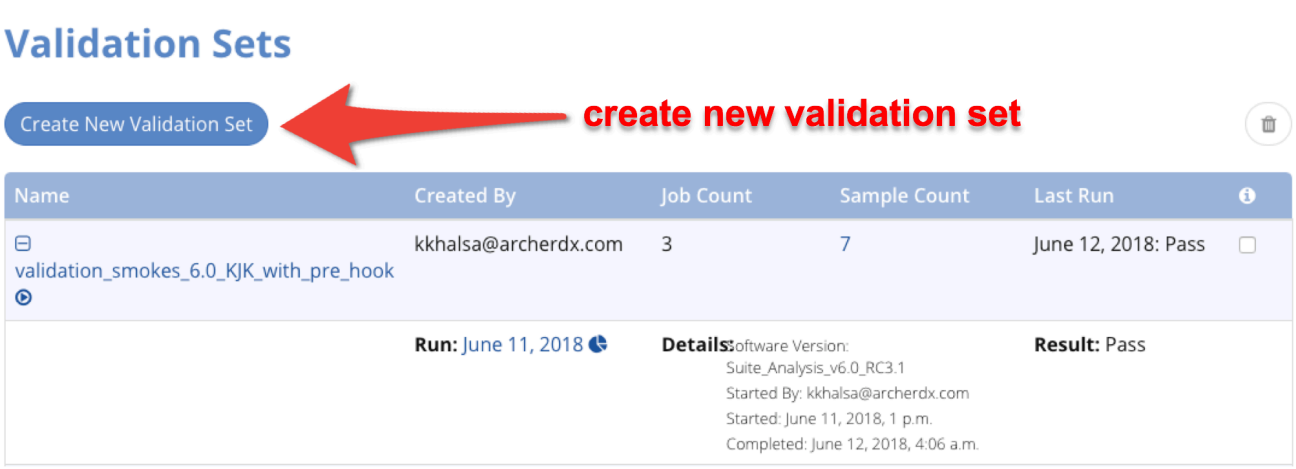
Marking Samples for Validation Set

Users have the ability to mark samples they wish to include in a validation set. This can be achieved by navigating to the "**Past Results**" page, selecting a specific job and locating the sample(s) you wish to incorporate into your validation set on the "**Sample Summery**" page. Once you locate a sample want to add, click on the ☰ icon and select the "**Mark as a Validation Sample**".

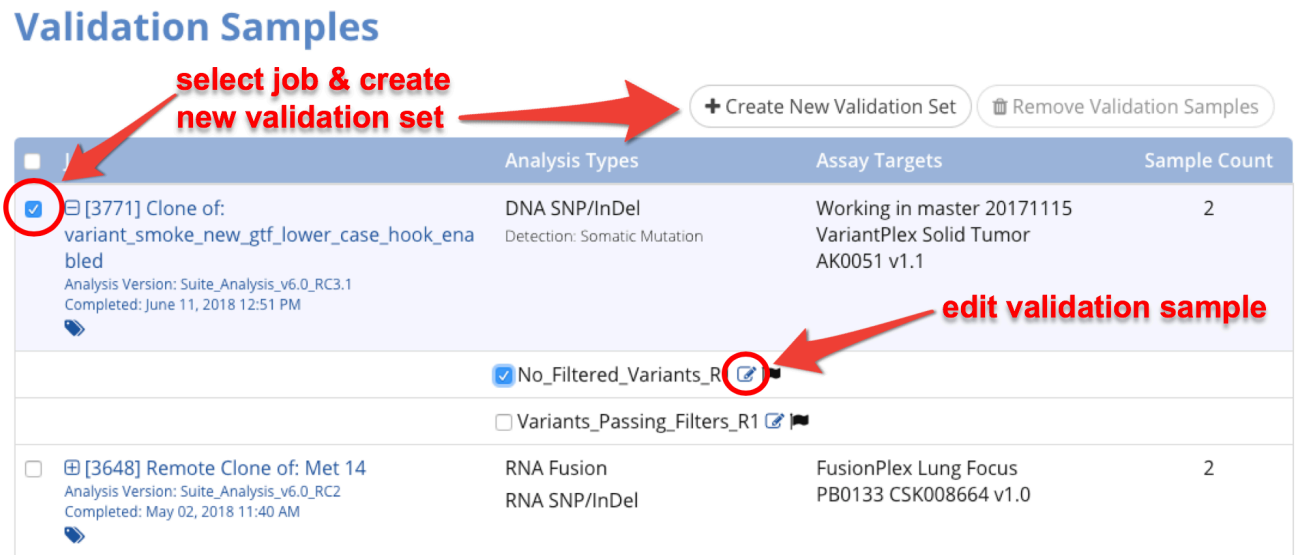


Once you have marked all the samples from a single job or multiple jobs you want to include in a validation set, you will then move on to creating a new validation set.

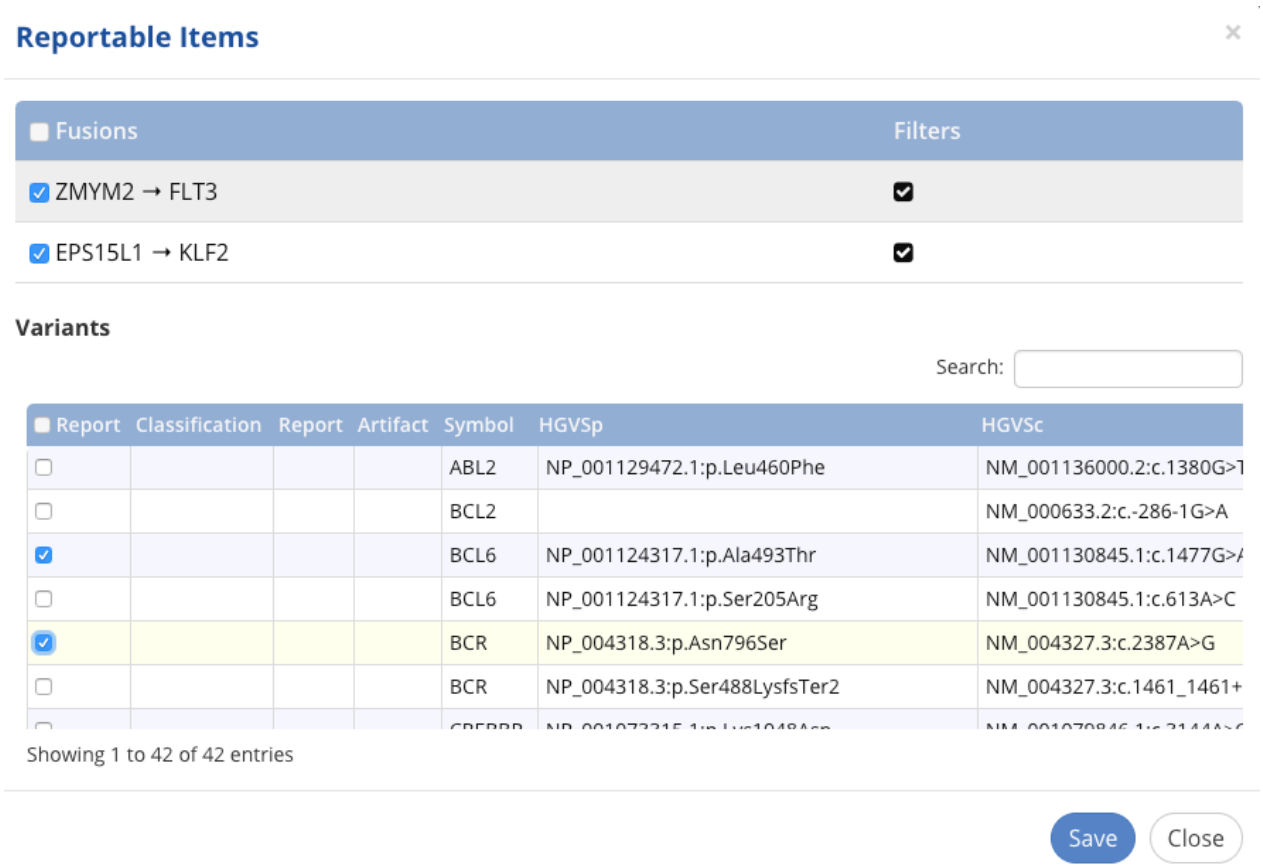
Creating/Removing a Validation Set



Once on the "**Validation Sets**" page, click on "**Create New Validation Set**", which will take you to the "**Validation Samples**" page.



Next, select all the jobs/samples you wish to include in the set and click on the **+ Create New Validation Set**" button. You will then be prompted to name your validation set and then select the "**Ok**" button. Users will then be directed back to the "**Validations Sets**" page where they will see their newly created validation set. Users have the ability to edit each validation sample to only report specific mutation types found within a sample. After clicking on the edit box icon, this will show a "Reportable Items" box where you can select the mutations types that you want part of the validation testing.



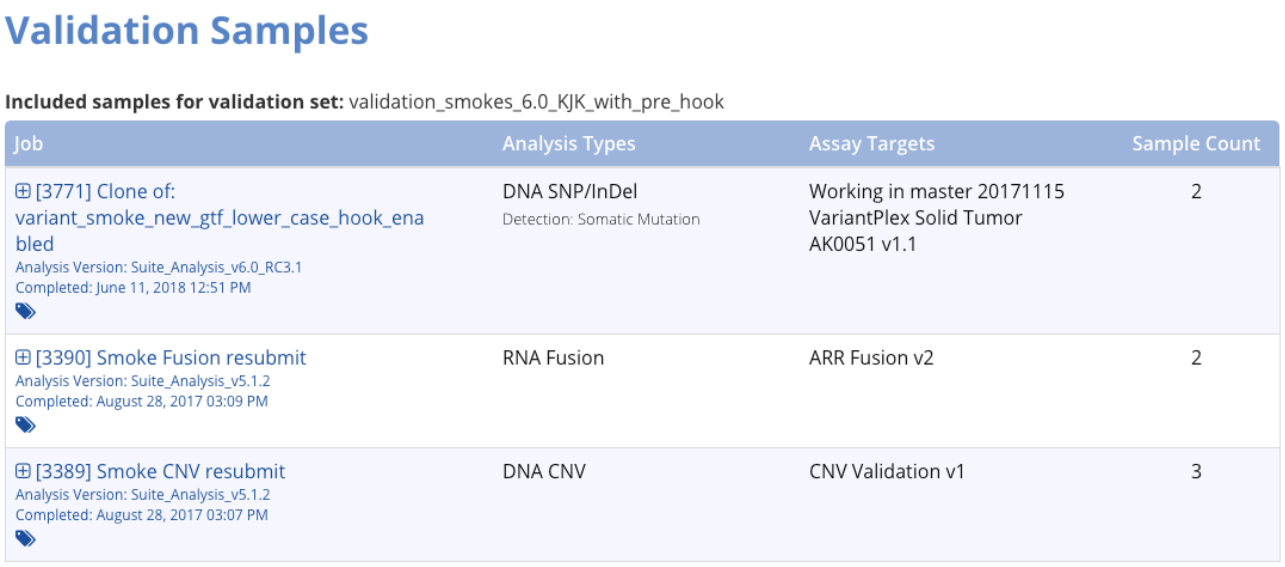
To remove a validation set(s) from the "**Validation Sets**" page, check the box next to each set you wish to delete and click the **icon** in the upper right part of the screen.

Running a Validation Set

Once you have created your new validation set, you will want to run the set to ensure the software is working properly. On the "**Validation Sets**" page, you will see the un-validated set listed with a next to the name of the validation set. Select this icon to begin running the validation process.

Validation Samples

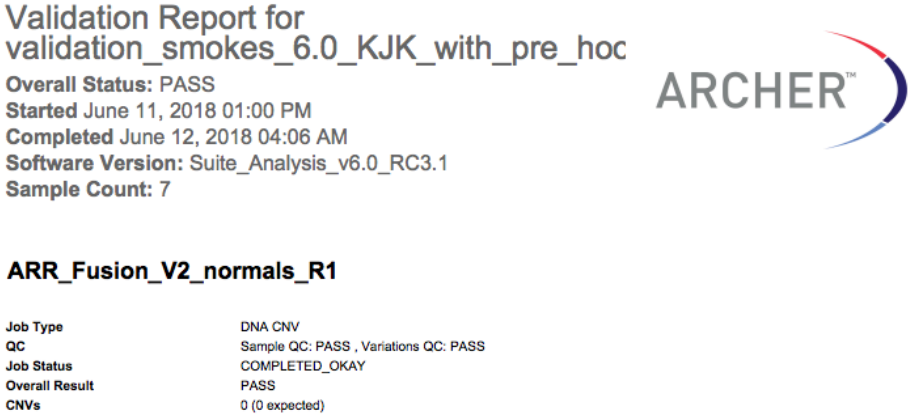
Users can view validation jobs/samples that are contained within a validation set. Click on the number of samples shown under the "**Sample Count**" column to access the jobs/samples on the "**Validation Samples**" page.



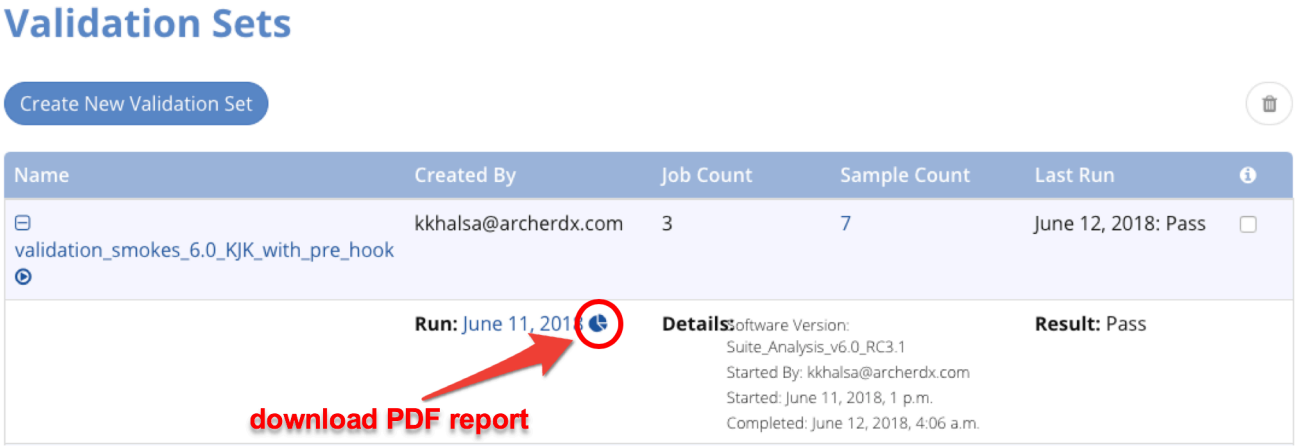
From here, users can view the details about the jobs/samples, analysis types, assay targets and sample count. Users can also add/remove samples from this this page if needed.

Validation Report

Regardless of a validation set passing/failing, a detailed validation report PDF is generated to provide the necessary information to the user about the status of the validation run. Details are generated for each sample that include name, job type, QC status, overall results and what was detected/expected.



To download the a Validation Report, navigate to the "**Validation Sets**" page, click the validation set, then click on the (pie chart icon ◔) next to the date the test was ran.



Run History

Validation set run history can be view on the "**Validation Sets**" page. This information will include details such as, name of the validation set, who created the set, job count, sample count and when it was last run.

Accessing and Interpreting Results

Key Assay Concepts

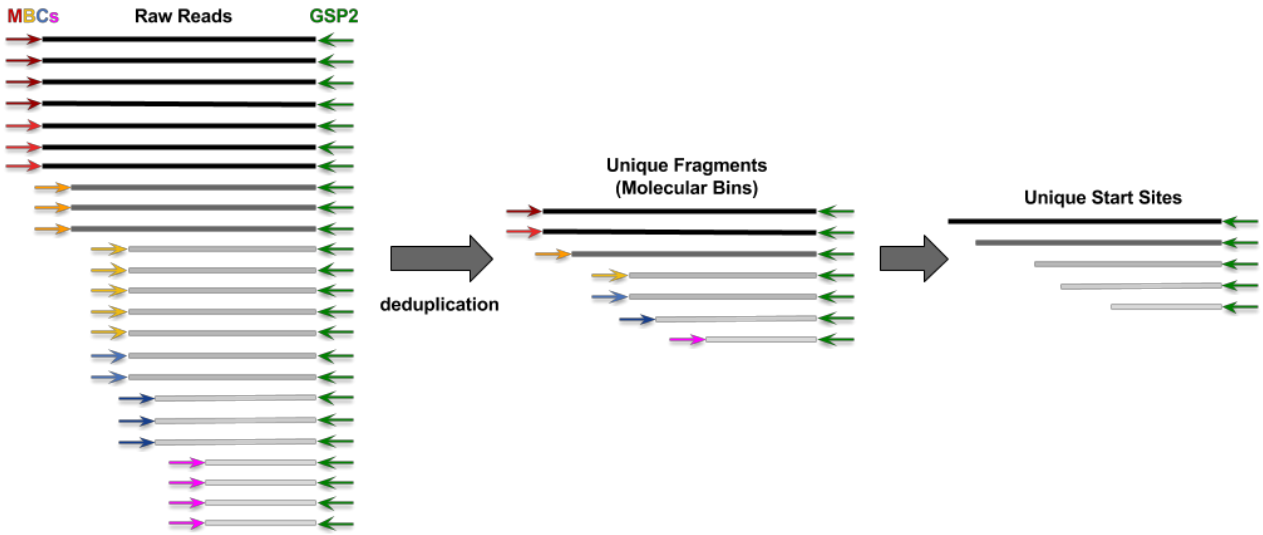
There are several key concepts that should be understood when working with Anchored Multiplex PCR (AMP) chemistry, in order to understand many of the terms and metrics below, as well as to properly interpret results from all Archer assays. These concepts are reviewed in the sections below.

Genome Reference

All Archer assays are designed for, and results presented relative to, hg19 (GRCh37).

Raw Reads, Unique Fragments, and Unique Start Sites

The graphic below illustrates how Analysis leverages the Molecular Barcodes (MBCs) of AMP in order to work backward from all reads reported from the sequencer (i.e. in the fastq files inputted at job setup), in order to deduce the original molecules that were extracted from your sample, and were inputted into the Archer target enrichment protocol:



Raw Reads (a.k.a. Total Fragments) – this is the total number of reads (or read pairs) reported from the sequencer.

Unique Fragments (a.k.a. Molecular Bins) – this is the total number of original fragments extracted from your sample, i.e., after fragmentation but before any PCR steps. These are determined by finding all raw reads (or read pairs) that share the same MBC, and deduplicating these into a single consensus sequence (note, low quality raw reads are filtered out before this step. See *Read Statistics* for further details).

Unique Start Sites – this is a subset of Unique Fragments, and represents the total number of unique fragment lengths extracted from your sample (since each fragment is anchored by a GSP2 on one side, length can only be affected by the random start site on the opposite end of the fragment). Unique start sites can manifest as multiple physical molecules, which is why there are typically more Unique Fragments than Unique Start Sites.

Unique Fragments tell you the true input quantity from your sample, and allow for accurate detection of allele fractions (AF). Unique Fragments and Unique Start Sites together provide a reliable measure of sample complexity, which can be correlated to assay sensitivity. These metrics also help guard potential false positives arising from PCR and sequencing error.

RNA vs. DNA vs. Ambiguous Reads

In RNA-based assays, it is likely that some fraction of reads originated from DNA, and it is important to differentiate which reads represent transcripts. To accomplish this, Unique Fragments that pass quality control (see Read Statistics for further details) are parsed further into three categories:

RNA Reads – those that span an exon-exon junction, as only transcripts would have such content.

DNA Reads – those that span an exon-intron junction, since RNA would rarely have such content.

Ambiguous Reads – those that are contained within a single exon (which makes it impossible to apply either of the rules above), or split reads that align to different genes that are not in Quiver.

The SNP/InDel pipeline (for both DNA and RNA) uses all reads from GSP2s marked SNV in the target GTF. The CNV Pipeline uses all reads from GSP2s marked CNV in the target GTF. (There is no reverse-transcription step in DNA assays, thus it is virtually impossible to get RNA reads from VariantPlex and Reveal assays).

The Immune Repertoire pipeline does not employ the above rules to differentiate RNA vs. DNA reads, and instead utilizes a third-party tool, MiXCR. See Immune Repertoire (RNA only) for more details.

Sample Tracking

We offer two sample tracking pipelines that enable users to predict sample sex and report a string of SNP genotype calls which are likely to be unique to each sample donor. The sex enumeration and string representing the genotype of all SNP_ID targeted variants can be used to troubleshoot sample swapping. The results are not intended to be used to assess low-level sample cross-contamination. The pipelines that yield the previously mentioned identifiers are referred to as SEX_ID and SNP_ID, and utilize primers that target the X and Y chromosomes and a series of highly polymorphic genomic positions respectively. Only Archer panels that contain these primers are capable of reporting this data. The pipeline is evoked via function flags in the GTF for these primers: function "SEX_ID" and function "SNP_ID".

Note, these primers target non-expressed DNA sequence (intronic or intergenic). As such, input samples must contain DNA, otherwise insufficient coverage will be generated from these primers. When using SEX_ID and SNP_ID with FusionPlex panels, do not use DNAse treated RNA as an input into library preparation.

SEX_ID

Some Archer panels contain SEX_ID primers that can be used to determine the sex of the donor. If a panel containing SEX_ID primers and the appropriate function flag in the GTF is used, there will be an output file that has "*molbar.trimmed.deduped.snp_id.txt" as a suffix. This file reports the sex of the donor based on data from two primers. The primers bind ambiguously to both X and Y chromosomes however the reads that extend off these primers generate sequence that unambiguously maps downstream from the primer This provides the ability to determine the chromosome of origin of the resulting fragments. There is a read depth requirement of 20x total unique reads (summed across the two primers) to make a call; if this is not met, the algorithm results in Unknown. If coverage on chrY (>20x unique reads) exists, the initial designation is male. If no coverage exists on chrY, the algorithm reports female. If there is coverage in Y, and also at least one heterozygous SNP call in X, the call is changed to Undetermined.

SNP_ID

Some Archer panels contain SNP_ID primers that can be used to report genotype across a series of highly polymorphic genomic positions in mostly intronic and intergenic regions. If a panel containing

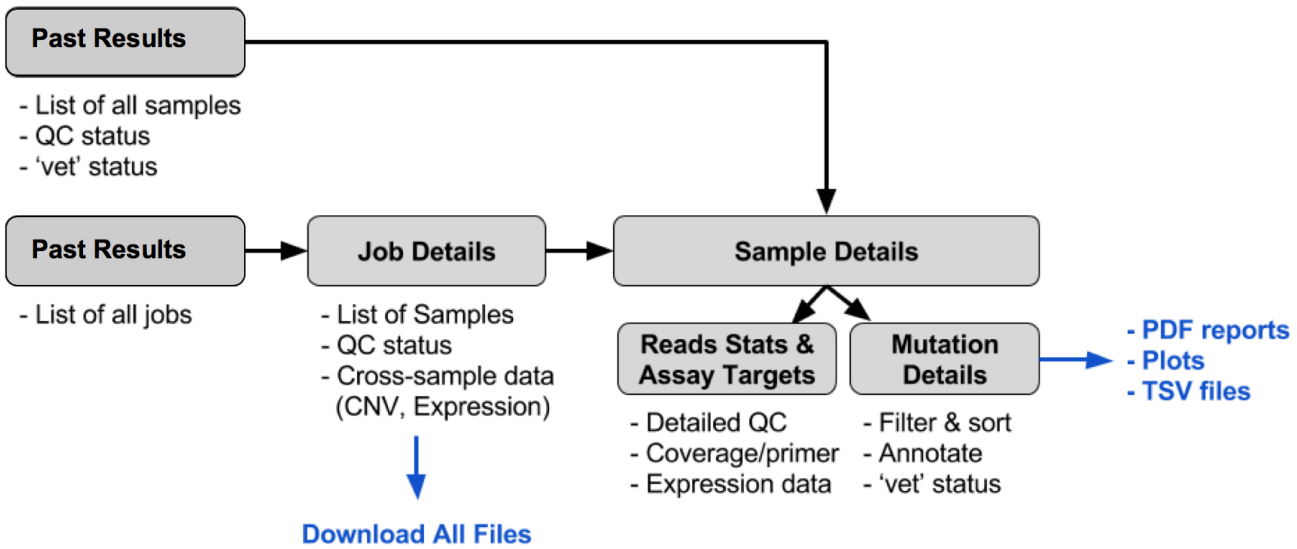
SNP_ID primers and the appropriate function flag in the GTF is used, there will be an output file that has "*molbar.trimmed.deduped.snp_id.txt" as a suffix, which contains the genotype report. Our in house targeted variant caller, VISION is used to determine the genotype of a SNP call across the targets and the following is reported for each target in the snp_id.txt file: chromosome, position, Ref, Alt, genotype, unique AF, unique RO and unique AO. If a the sum unique AOs and unique ROs is less than 20 the genotype assignment is deferred. The genotype call for each position is classified as one of 4 types: HOMOZYGOUS REF, HETEROZYGOUS, HOMOZYGOUS ALT, or UNDETERMINED. These calls are abbreviated as "r", "h", "A" or ".", respectively. The first line of the snp_id.txt file contains the series of abbreviations as a sample ID string.

Control Genes (FusionPlex Only)

Since RNA expression can vary widely based on a number of biological factors, 8 control GSP2s targeting 4 genes are included in FusionPlex assays. These 4 genes, CHMP2A, GPI, RAB7A, and VCP, are known to be consistently moderately expressed across most tissue types, and thus collectively serve as a reliable (1) indicator of overall RNA quality and content in the sample, and (2) benchmark for normalization in the RNA Expression pipeline. See *Sample QC Fusions QC* and *Relative Expression* for further details on these points.

General Results Review & Interpretation Workflow

The GUI is designed to present all information needed to complete secondary analysis (i.e. true/false positive determination), as well as some amount of tertiary analysis (e.g. using Clinvar annotations), but all data is accessible for download and custom processing. There are two primary ways to access results of completed jobs in Archer Analysis, by sample (**Recent Samples**) and by job (**Past Results**). The diagram below illustrates the basic job/sample/mutation review and sign-out workflow. There are a number of points throughout the workflow to extract results, either for final reporting (PDFs and image files) or for further data processing (BAMs, raw text files, etc.):

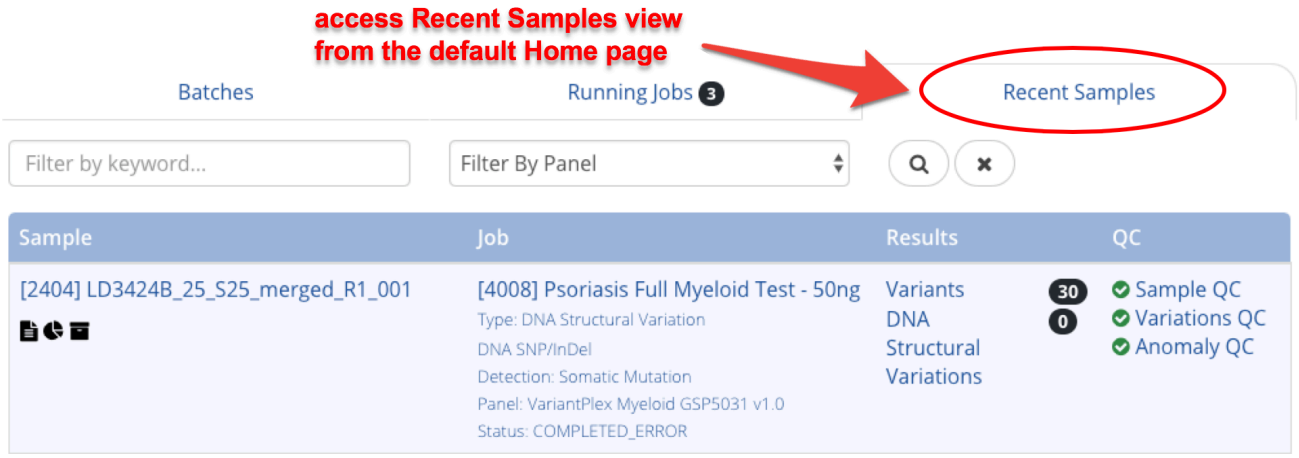


Batches

The **Batches** view is intended as a quick way to screen through a daily caseload, and is a means to access all batches that you either own, or have been assigned to you. Only batches that have not been completely reviewed are listed here. The view is accessed by selecting the "**Batches**" tab from the default Home page (which can be changed to this view, see *Miscellaneous System Settings* for further details):

Recent Samples

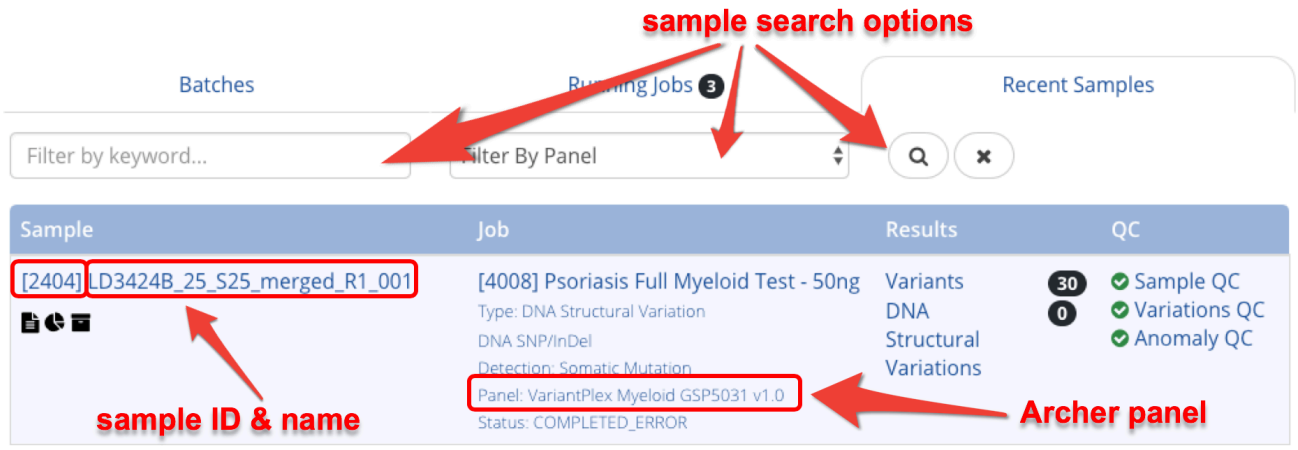
The **Recent Samples** view is intended as a quick way to screen through a daily caseload, and is a means to access all samples in the system. The view is accessed by selecting the "**Recent Samples**" tab from the default Home page (which can be changed to this view, see Miscellaneous System Settings for further details):



By default, the samples are listed according to:

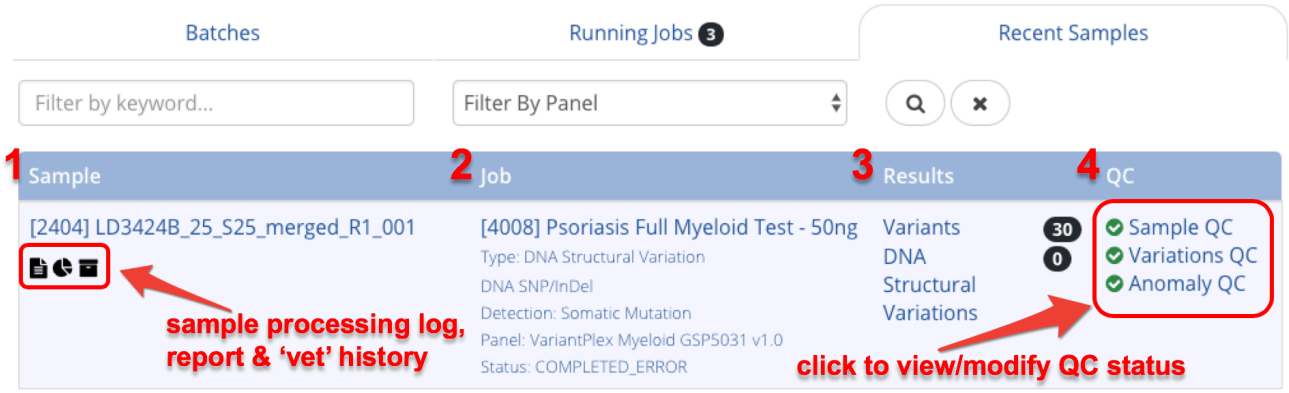
1. Most recent completed job
2. Then by the order the sample was processed (i.e. ascending sample ID)

Samples can also be found by searching by keyword (e.g. sample name or ID), by Archer Panel, and/or by 'Vet Status' (i.e. Not Reviewed, Reviewed, Signed Out, Other):



Sample Summary

The **Recent Samples** page provides a summary of key sample data, as well as links to take action (i.e. review, sign out) on each sample:



1. Click the sample ID/name under the ***Sample*** column to jump to the **Sample Details** page and get an in-depth view of all results for the sample (see Sample Details Page for further details).
2. Click the job ID/name under the ***Job*** column to jump to the **Sample Summary (Job Details)** page of the job the sample was run in (see Job Details Page (Sample Summary) for further details).
3. Click on any of the links under the ***Results*** column in order to jump right to the pipeline-specific results (e.g. Fusions, Variants) within the **Sample Details** page. See Pipeline-Specific Information for further details.
4. Click on either the name of the QC type (e.g. Fusion QC) or the pass/fail icon (✔ / ✖) to view the **Sample QC History** popup. See QC Results for further details.

These links are also available in the sample summary of the **Recent Samples** page:

- Download the sample log file (#.log.stderr.txt) by clicking the 🗎 icon.
- Generate a PDF report by clicking the ◔ icon.

- View and edit sample 'vet' history by clicking the 💬 icon.

QC Results

There are four Quality Control (QC) metrics reported at the sample level to determine if there is any potential issue with the sample related to the extraction, library prep, and/or sequencing: Sample QC, Anomaly QC, Fusions QC and Variations QC:

Sample QC

This is the overall QC result for a sample. The value is the aggregation of the individual metric QC results. Please note that Immune Repertoire do not have a Sample QC metric.

Anomaly QC

The Anomaly QC metric will report pass/fail only if the Structural Variations pipeline is selected for. It is defined as the average number of unique start sites (from DNA and ambiguous reads) calculated per GSP2 across the entire panel. (See Key Assay Concepts for explanations of unique start sites and DNA/ambiguous reads).

Fusions QC

The Fusions QC metric will report pass/fail only if the Fusions/Isoforms pipeline is selected for FusionPlex assays (i.e. "RNA Fusion" Analysis Type). It is defined as the average number of unique start sites (from RNA reads) calculated per control GSP2 (see Key Assay Concepts for explanations of unique start sites, RNA reads, and control GSP2s).

This metric is considered a reliable measure of sample complexity (i.e. information content) of the sequencing library, with low values indicating that either (1) sample input was insufficient quality and/or quantity, or (2) there was some issue with extraction, library prep, and/or sequencing.

This metric is user-configurable, and should be calibrated per your lab's specific requirements and operating environment. See *General Analysis Settings (MIN_AVERAGE_UNIQUE_DNA_AND_AMBIG_START_SITES_PER_GSP2)* for further details.

The PreSeq™ RNA QC Assay is correlated to this Analysis QC metric, and thus provides a means of rescuing samples early in library preparation from further processing. Consult the FusionPlex protocol for your assay for further details on using PreSeq.

Variations QC

The Variations QC metric will report pass/fail if the SNP/InDel and/or CNV pipelines are selected. It is defined as the average number of unique start sites (from DNA and ambiguous reads) calculated per GSP2 across the entire panel. (See *Key Assay Concepts* for explanations of unique start sites and DNA/ambiguous reads).

This metric is considered a reliable measure of sample complexity (i.e. information content) of the sequencing library, with low values indicating that either (1) sample input was insufficient quality and/or quantity, or (2) there was some issue with extraction, library prep, and/or sequencing.

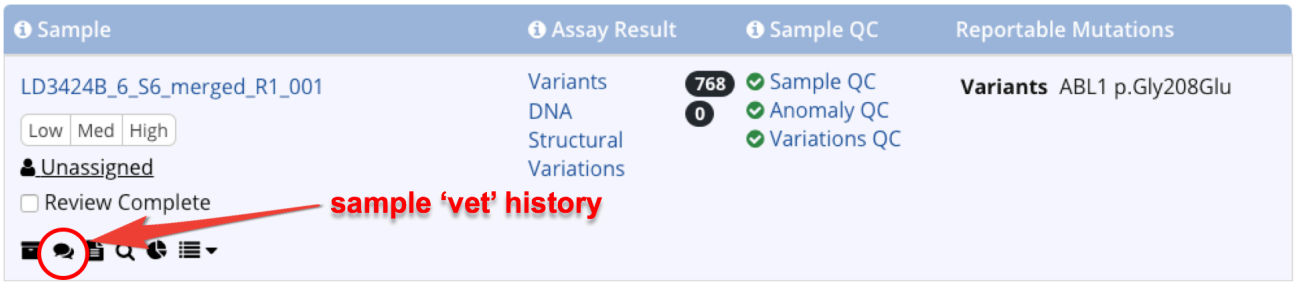
This metric is user-configurable, and should be calibrated per your lab's specific requirements and operating environment. See *General Analysis Settings (MIN_AVERAGE_UNIQUE_DNA_AND_AMBIG_START_SITES_PER_GSP2)* for further details.

Viewing and Overriding QC Pass/Fail Threshold

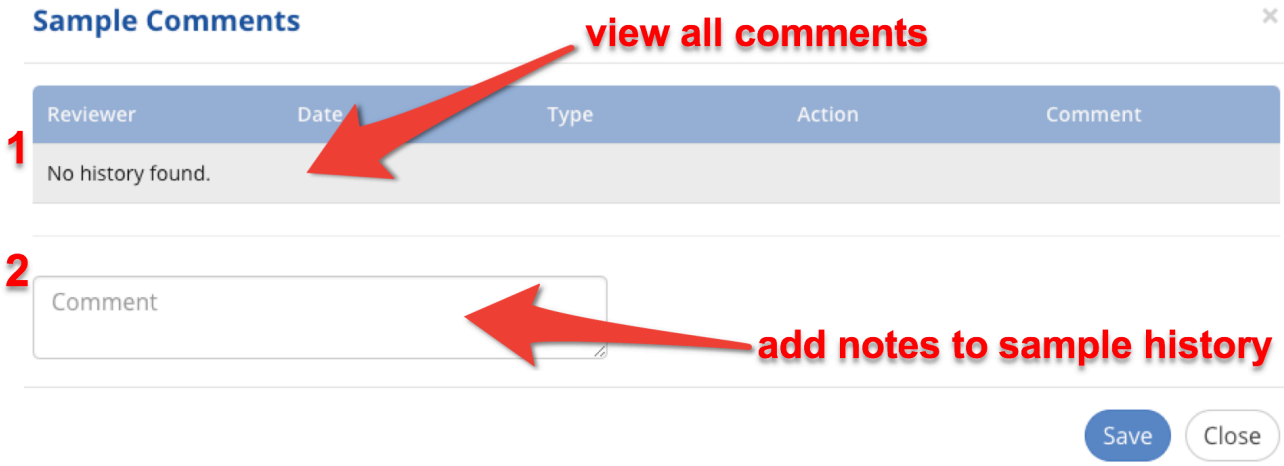
Click on either the name of the QC type (e.g. Fusion QC) or the pass/fail icon (✔ / ✖) to view the **Sample QC History** popup, which shows not only the QC metric for the sample, but also what the threshold is (as defined by User Settings at the time of job setup; see *General Analysis Settings* for further details). This popup also allows for manual override of the pass/fail call, along with commenting on the sample – all actions are recorded and viewable in the History section at the top:

Results Review and Reporting

The core intent of the **Batches** and **Recent Samples** page is to review and report on each sample. A complete history of any comments and change in a sample's 'view comments' status is tracked and modified by clicking the icon:



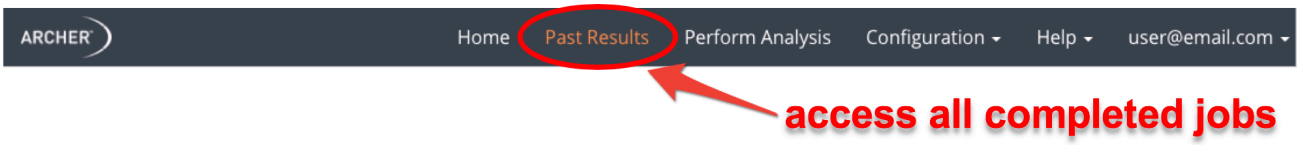
This will bring up the **Sample History** popup:



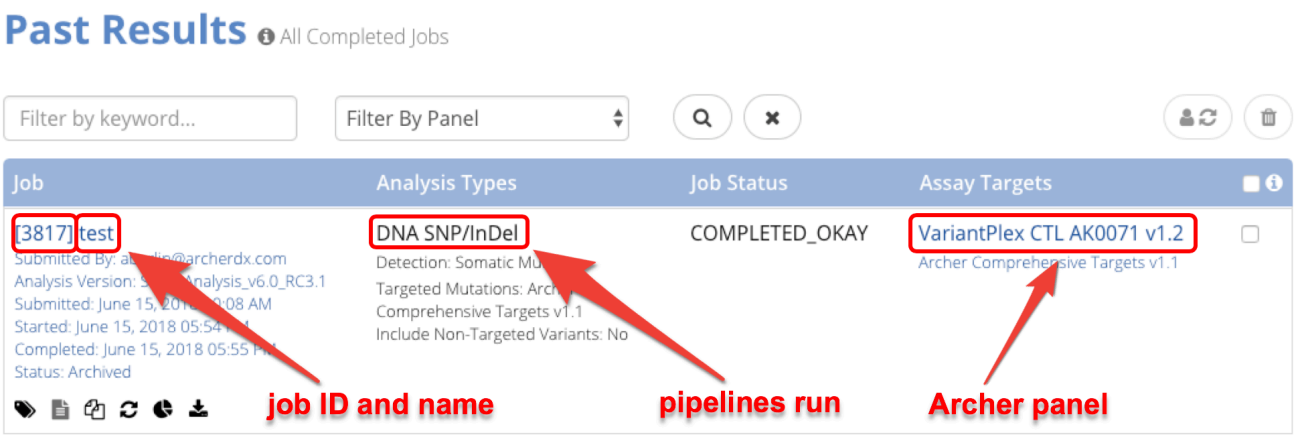
1. View entire 'vet' history. This includes change in status and/or comments added. Previous additions cannot be modified.
2. Add a comment to the sample history. This can be done independently of 'vet' status change.

Past Results

Access the **Past Results** page via the main menu:



The **Past Results** page also shows a list of all completed jobs, including any jobs run by other users that the current user has permission to view (see Managing Users and Groups for further details) Specific jobs can be located by typing a query into the search box, or by filtering by any GTF files currently loaded into the system that have associated jobs:

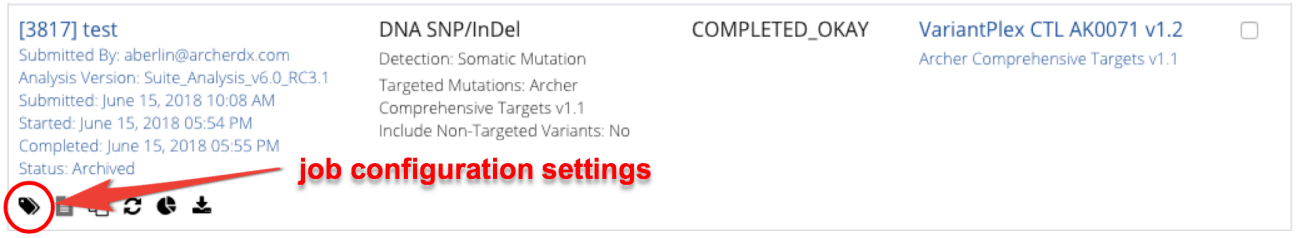


The job list includes descriptive information about each job:

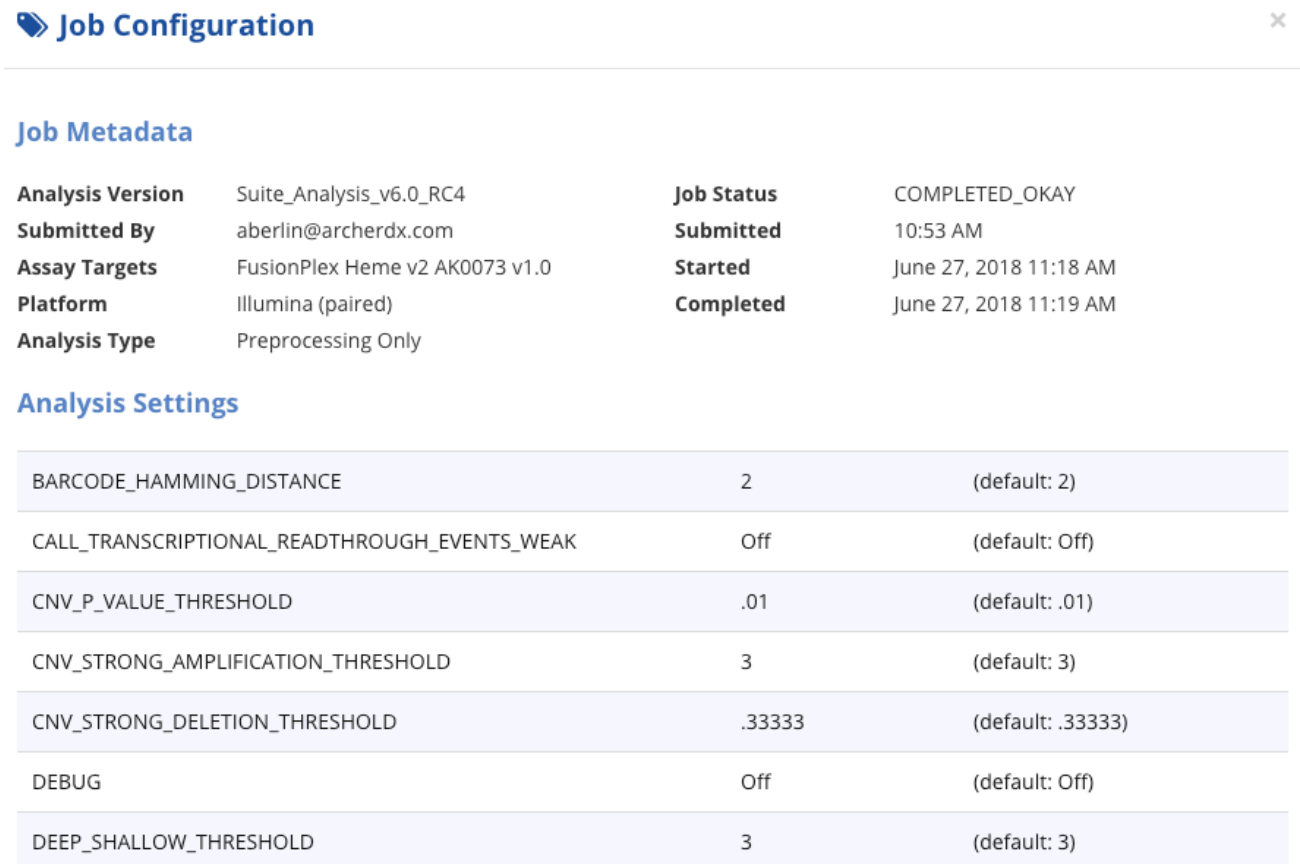
- **Job** - Job ID, job name, and timestamps for when each job was submitted, started, and completed. This column also contains several clickable icons, the functions of which are described in the following sections.
- **Job Details** - The type of analyses that were chosen when setting up the job.
- **Job Status** - One of “NEW”, “QUEUED”, “RUNNING”, “HALTED”, “COMPLETED_ERROR” and “COMPLETED_OKAY”.
- **Assay Targets** - The panel (GTF file) that was chosen when setting up the job. If a Targeted Mutations file was chosen, it is also shown here.
- **Platform** - The sequencing technology that was chosen when setting up the analysis.

Reviewing Settings of a Job

Click on the Job Configuration icon (🏷) to see what parameter values (user settings) were used for the processing of the job:



The default values of each parameter are also shown. If any settings differ from the default, they will be highlighted (see Identify Which Settings are Not Default for further details):

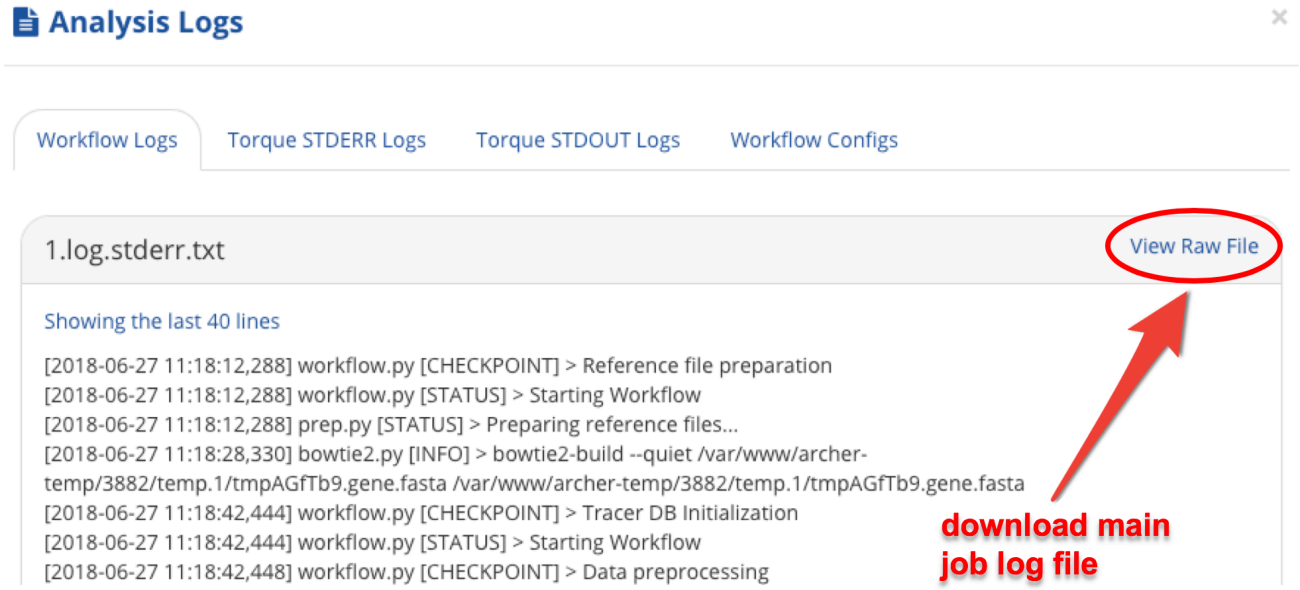


Reviewing (Error) Logs

After a run has completed, the status of the job should be “COMPLETED_OKAY”. If this is not the case, review the error logs to determine the cause of the failure. To review the error log, select the file icon (🗎) below the job details for the run:

A tabbed page will show the Workflow Logs, the Torque STDERR (Standard Error) log and the Torque STDOUT (Standard Out) log. Review the Torque STDOUT log file first, since this could also contain error messages.

In addition, click the "View Raw File" link to download the main log file to the local computer:

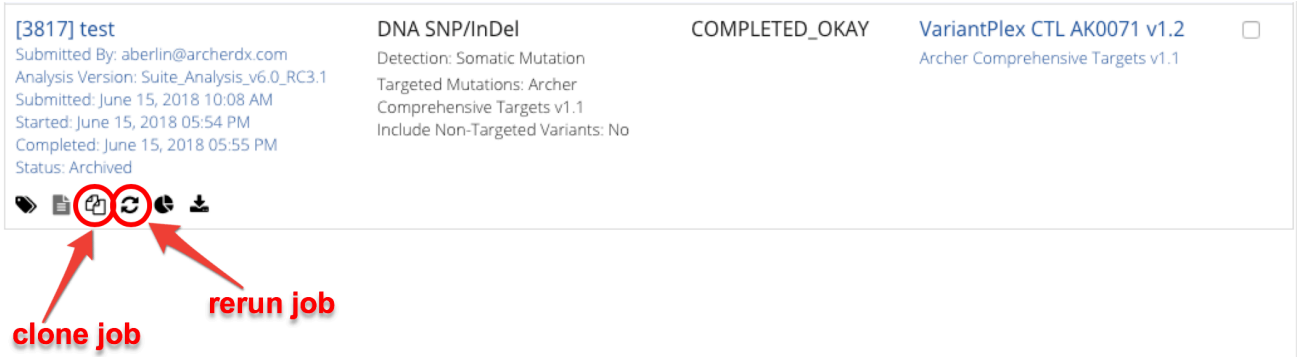


Contact tech@archerdx.com for further assistance and troubleshooting.

Rerun or Clone a Job

If the results of a run are not satisfactory, it may be useful to re-run an analysis with different analysis settings. This can be accomplished by either making a clone of the original analysis or by re-running the original analysis. The difference between the two options is that a clone will be a copy of the original run, leaving the old run intact, while re-running an analysis will remove the old results and replace it with the new (i.e., all existing data except the fastqs will be deleted for the job).

To clone an analysis, select the “Clone Job” icon (), and to re-run an analysis, select the “Rerun Analysis” icon ():

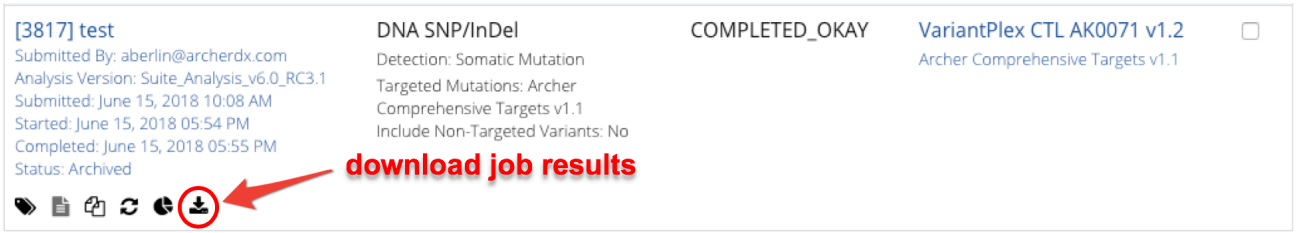


View Overview of Results for the Entire Job (Reports)

To generate a PDF report showing the Fusion QC results and a summary of the analysis results for each sample, select the "Job Reports" icon ().

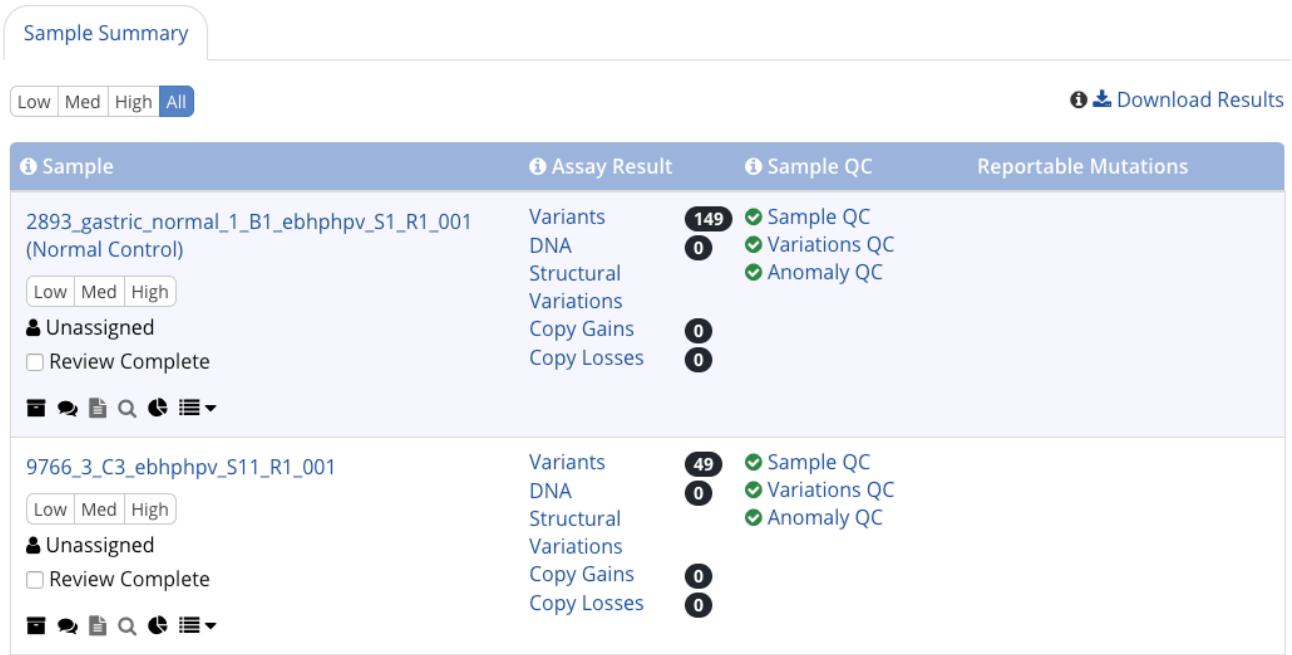
Download Results for the Entire Job

To download analysis results for the job, select the "Download Results" icon (📥).



Job Details Page (Sample Summary)

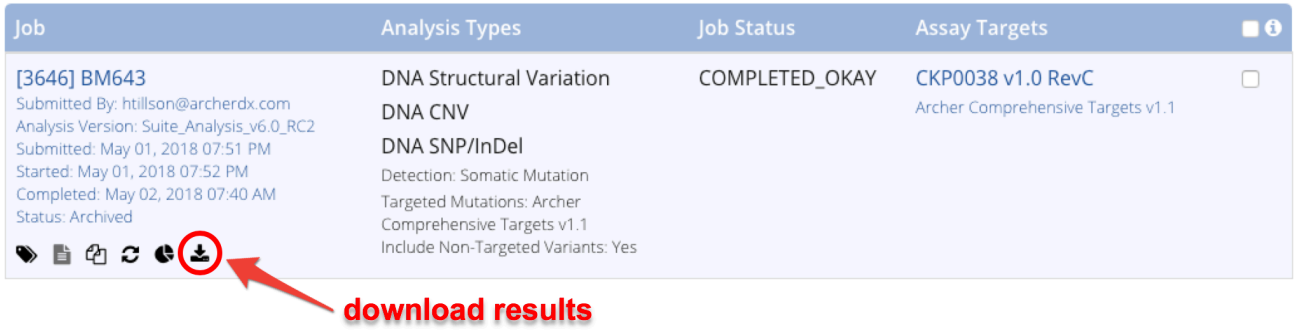
From the **Running Jobs**, **Past Results**, or **Recent Samples** page, click on the job ID/name link to see the Job Details page. The page will show a list of all samples in the jobs, with various clickable icons below them, as well as the “Assay Result” and “QC Result” for each sample. The specific “Assay Result” categories that are shown will depend on how the job was set up (see *Starting a Job: Perform Analysis Page* for further details):



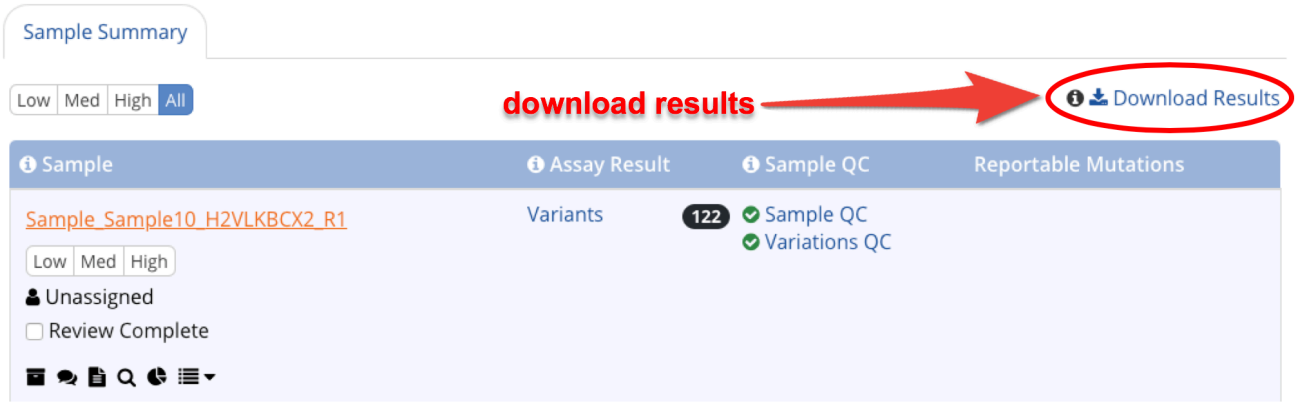
User Defined File Download List

Choose Files to Download

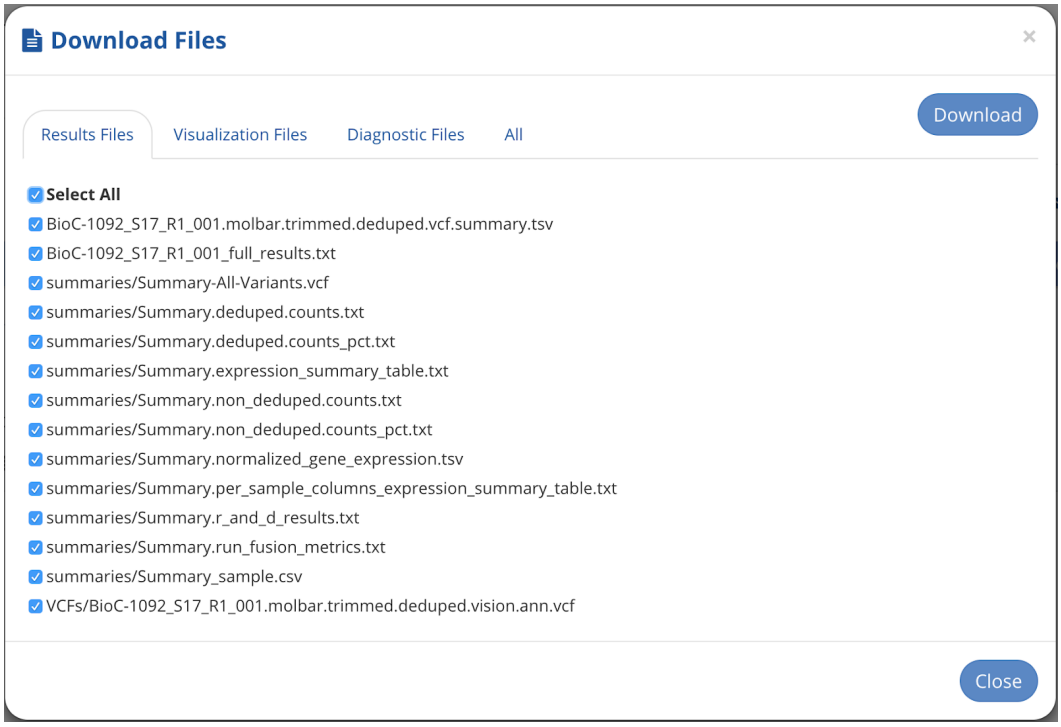
To download files for the job select the "**Download Results**" icon (). This will open up a pop up view that will allow manual selection of specific (or all) files that can then be downloaded as a single ZIP. For more information see the User Defined File Download List section.



All files, including summaries, etc., that were generated for the job can be downloaded as specified by the user, by clicking the "Download Results" icon ().

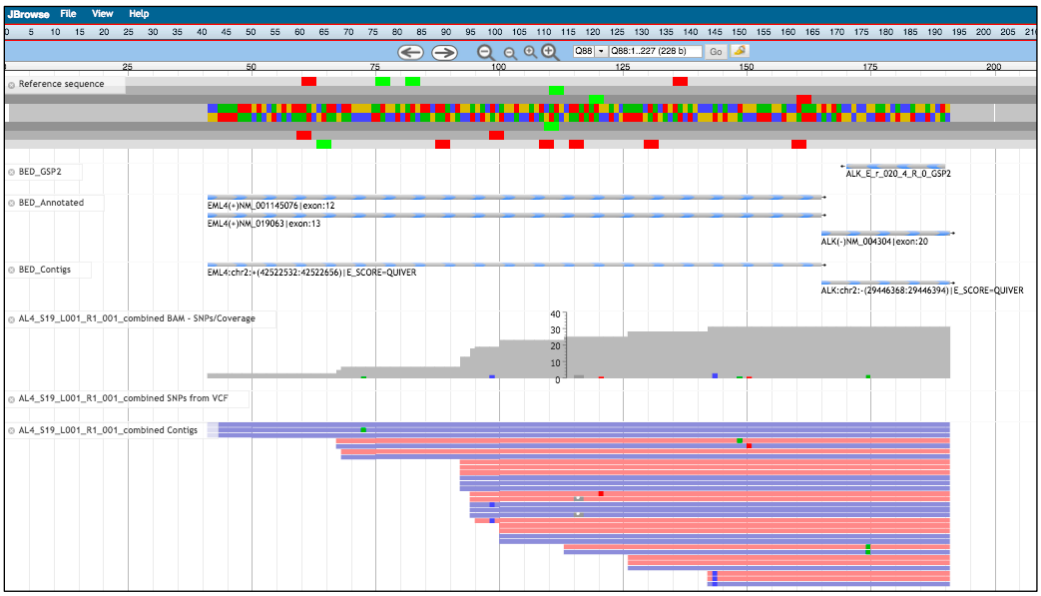


From this view all files produced by the job can be browsed and selected, pre organized into tabs with commonly downloaded files. After selecting at least one file, the selection can then be downloaded as a single ZIP file.



Visualize Sample Reads in JBrowse

From the **Sample Summary** page, the "Visualize Sample Data" link (🔍) will open a window with the JBrowse genome browser showing all the reads aligned to the human genome (hg19) that formed the basis for the Gene fusion and/or Mutation analysis:



The genome browser shows a number of different tracks:

***Reference Sequence* –** A track showing the hg19 reference sequence with both strands and the 6-frame translation as well.

***refGene* –** A track showing the genes and each of their transcripts from the RefSeq database. Right-click on the transcript to see its name and follow the link for a search of the transcript at the NCBI website.

***Target Region GSP2* –** A track showing the position and direction of the gene specific primers.

***Cosmic* –** A track showing the location and identifiers from the COSMIC mutation database.

Right-click on the feature to follow the link of the identifier to the Sanger website

***Merged BAM Coverage* –** Both Read 1 and Read 2 are combined into a single BAM file.

***Preprocessed BAM Coverage* –** BAM coverage after filtering, InDel realignment and deduplication.

The search box at the top can be used to navigate to a specific genomic location directly by typing in the location OR can be used to enter a gene name.

For more information on usage of JBrowse, see the manual for JBrowse via the help links in the JBrowse window.

Sample Details Page

Below each sample name of the **Job Details** page is a link that will take you to the **Sample Details** page for that sample. Clicking on this link will show the summary results pages with more details. This page displays the detailed results of the analysis and QC metrics. A left-hand side menu provides all links to all results, review/report status, and any other options relevant to the sample:

Read Statistics

The reads statistics page contains basic metrics for the sample library, such as mapping percentages, on target percentages, and DNA/RNA statistics:

Molecular Barcode Statistics

Total Fragments	Fragments with Complete Adapter	Number of Reads After Trimming Adapters
2,183,902	2,114,980	1,768,190

Export Data (tsv)

Read Statistics

Type	Total Fragments (# / %)	Mapped (# / %)	Pass Alignment Filter (%)	On Target (%)
All Fragments	1,768,190 / 100.0	1,767,403 / 100.0	100.0	85.2
Unique Fragments	1,087,293 / 61.5	1,086,599 / 99.9	99.9	81.4

Export Data (tsv)

DNA/RNA Statistics

Type	DNA Reads (# / %)	RNA Reads (# / %)	Ambiguous Reads (# / %)
All Fragments	1,506,100 / 100.0	0 / 0.0	0 / 0.0
Molecular Bins	884,921 / 100.0	0 / 0.0	0 / 0.0
Average Molecular Bins per GSP2	1,340.79	0	0
Unique Start Sites	55,370 / 100.0	0 / 0.0	0 / 0.0
Average Unique Start Sites per GSP2	96.20	0	0
Average Unique Start Sites per GSP2 Control	0	0	0

Export Data (tsv)

QC Statistics

Avg. Unique DNA And Ambiguous Start Sites Per GSP2
180.5

Miscellaneous Statistics

On Target Deduplication Ratio
5.17:1

Export Data (tsv)

DNA/RNA Fragment Lengths

DNA Median Fragment Length	DNA Mean Fragment Length	RNA Median Fragment Length	RNA Mean Fragment Length
204.0	222.4	No Fragments	No Fragments

Export Data (tsv)

Most of the metrics contain help texts (mouse-over the icons) and a full description of the fields can be found in Read Stats & QC Metrics.

Assay Targets

The assay targets page contains detailed coverage information for each GSP2 in the panel. It contains information about the number of unique fragments based on the number of unique molecular barcodes, as well as the number of unique start sites, and lastly raw (non-deduplicated) reads. The data are also separated by the type of molecule from which the fragments originate (DNA, RNA or AMBIGUOUS). In addition to the coverage for the actual primers, coverage data is also provided for those reads that fall anywhere inside the gene locus. Those targets can be recognized by the “NEAR” tag:

Assay Targets

Unique Molecular Bins | Unique Start Sites | DNA | RNA | Ambiguous | Total

Target	Fragments (# / %)	RNA Fragments (# / %)	DNA Fragments (# / %)	Ambiguous Fragments (# / %)
AKT1_chr14_105246503_25_+_A1_GSP2	195 / 0.036	0 / 0.000	195 / 0.036	0 / 0.000
AKT1_chr14_105246579_20_-_A1_GSP2	202 / 0.037	0 / 0.000	202 / 0.037	0 / 0.000
ALK_chr2_29432634_27_+_A1_GSP2	247 / 0.046	0 / 0.000	247 / 0.046	0 / 0.000
ALK_chr2_29432692_21_-_A1_GSP2	233 / 0.043	0 / 0.000	233 / 0.043	0 / 0.000
ALK_chr2_29443539_24_+_A1_GSP2	216 / 0.040	0 / 0.000	216 / 0.040	0 / 0.000
ALK_chr2_29443634_21_-_A1_GSP2	245 / 0.045	0 / 0.000	245 / 0.045	0 / 0.000
ALK_chr2_29443665_25_+_A1_GSP2	266 / 0.049	0 / 0.000	266 / 0.049	0 / 0.000

The Assay Targets page is itself divided into seven different sections. The top section of the page is dedicated to Controls (for RNA assays; see Sample Tracking)

We offer two sample tracking pipelines that enable users to predict sample sex and report a string of SNP genotype calls which are likely to be unique to each sample donor. The sex enumeration and string representing the genotype of all SNP_ID targeted variants can be used to troubleshoot sample swapping. The results are not intended to be used to assess low-level sample cross-contamination. The pipelines that yield the previously mentioned identifiers are referred to as SEX_ID and SNP_ID, and utilize primers that target the X and Y chromosomes and a series of highly polymorphic genomic positions respectively. Only Archer panels that contain these primers are capable of reporting this data. The pipeline is evoked via function flags in the GTF for these primers: function "SEX_ID" and function "SNP_ID".

Note, these primers target non-expressed DNA sequence (intronic or intergenic). As such, input samples must contain DNA, otherwise insufficient coverage will be generated from these primers. When using SEX_ID and SNP_ID with FusionPlex panels, do not use DNAse treated RNA as an input into library preparation.

SEX_ID

Some Archer panels contain SEX_ID primers that can be used to determine the sex of the donor. If a panel containing SEX_ID primers and the appropriate function flag in the GTF is used, there will be an output file that has "*molbar.trimmed.deduped.snp_id.txt" as a suffix. This file reports the sex of the donor based on data from two primers. The primers bind ambiguously to both X and Y chromosomes however the reads that extend off these primers generate sequence that unambiguously maps downstream from the primer This provides the ability to determine the chromosome of origin of the resulting fragments. There is a read depth requirement of 20x total unique reads (summed across the two primers) to make a call; if this is not met, the algorithm results in Unknown. If coverage on chrY (>20x unique reads) exists, the initial designation is male. If no coverage exists on chrY, the algorithm reports female. If there is coverage in Y, and also at least one heterozygous SNP call in X, the call is changed to Undetermined.

SNP_ID

Some Archer panels contain SNP_ID primers that can be used to report genotype across a series of highly polymorphic genomic positions in mostly intronic and intergenic regions. If a panel containing SNP_ID primers and the appropriate function flag in the GTF is used, there will be an output file that has "*molbar.trimmed.deduped.snp_id.txt" as a suffix, which contains the genotype report. Our in house targeted variant caller, VISION is used to determine the genotype of a SNP call across the targets and the following is reported for each target in the snp_id.txt file: chromosome, position, Ref, Alt, genotype, unique AF, unique RO and unique AO. If a the sum unique AOs and unique ROs is less than 20 the genotype assignment is deferred. The genotype call for each position is classified as one of 4 types: HOMOZYGOUS REF, HETEROZYGOUS, HOMOZYGOUS ALT, or UNDETERMINED. These calls are abbreviated as "r", "h", "A" or ".", respectively. The first line of the snp_id.txt file contains the series of abbreviations as a sample ID string.

(See Control Genes (FusionPlex Only) for further details), and the section below is for the actual assay targets. Across all sections described below, data is sorted into Total, RNA, DNA, and Ambiguous reads (see RNA vs. DNA vs. Ambiguous Reads for further details):

- **Unique Molecular Bins** – defined as the total unique reads (i.e. de-duplicated reads).
- **Unique Start Sites** – defined as the number of start sites occupied by at least one read.
- **Raw Alignment** – represents the coverage data BEFORE de-duplication.
- **DNA** – represents the data for DNA reads only and further separated by DNA reads that have a unique molecular bin and those DNA reads that have unique start sites.
- **RNA** – represents the data for RNA reads only and further separated by RNA reads that have a unique molecular bin and those RNA reads that have unique start sites.
- **Ambiguous** – represents the data for ambiguous reads only and further separated by ambiguous reads that have a unique molecular bin and those ambiguous reads that have unique start sites.
- **Total** – represents the data for all reads combined (RNA, DNA and Ambiguous reads) and further separated by reads that have a unique molecular bin and those reads that have unique start sites.

Pipeline-Specific Tabs

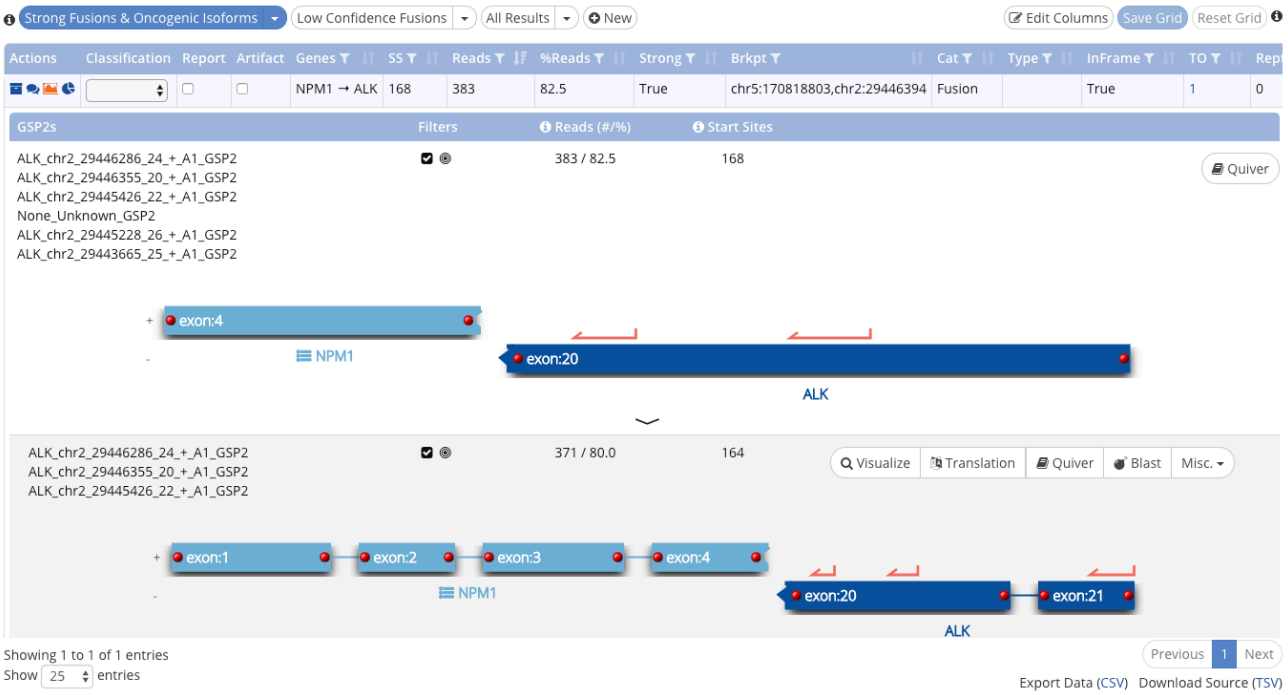
The additional tabs found in the **Sample Details** page show information for each specific pipeline selected at job setup. See Pipeline-Specific Information for further information on the info found in these tabs.

Pipeline-Specific Information

Fusion/Isoforms (RNA Only)

RNA Fusion pipeline will detect gene fusion events by annotating the de-novo assembled RNA reads with BLAST. Archer-specific filtering logic is used to reduce misalignments and false positives. SNP and InDel analysis can also be selected to include the detection of small variations with respect to the human hg19 reference.

The **FusionViewer** displays results of candidates of any Fusions or Isoforms from this pipeline:



The initial displayed fusion is aggregated by genomic breakpoint, showing where the GSP2 primer(s) are positioned on the exon. This fusion can be viewed by supporting isoforms. The aggregated visualization highlights the breakpoint by showing the 2 contigs adjacent to the breakpoint and lists all the primer(s) that support the fusion. Additional details are shown in the per fusion consensus visualizations listed under the aggregated fusion.

The following links are available in the **FusionViewer**:

Visualize

Opens a new window with the reads supporting the gene fusion

Blast

Opens a new window with the results of a BLAST search

Translation

Opens a new window showing the results of the protein translation prediction of the gene fusion product

A link on this page, marked Blast will open a new window with the results of a BLAST search against the human *protein* database.

Key Metrics for Strong and Low Confidence Fusion/Isoform Candidates

Reads (#/%) – The number and percentage of unique reads supporting this gene fusion based on the molecular barcode. Only reads spanning the breakpoint are considered to support the fusion. Paired reads where both reads completely cover only one of the genes are NOT considered as supporting a gene fusion. The percentage is calculated in reference to the total number of reads covering this target, including wild-type transcripts. The read must extend 5 bp past GSP2 in order for the fusion to be called on the GSP2 side & this read must read 30 bp into the fusion partner for this to be called.

Start Sites – The number of unique reads supporting this gene fusion based on the unique start sites. Only reads spanning the breakpoint are considered to support the fusion. Paired reads where both reads completely cover only one of the genes are NOT considered as supporting a gene fusion.

Breakpoint – The hg19 chromosomal breakpoint locations for the fusion as deduced from the RNA. This does NOT represent the exact breakpoint at the DNA level. (Coordinates at the start of chr1 are placeholders for unaligned bases).

Strong and Low Confidence Criteria

Gene fusions, oncogenic isoforms, wild type isoforms, and novel isoforms are separated into two categories:

1. Those with strong support for the call
2. Those with either weak support for the call, or with characteristics indicative of a false positive call

The software provides a Strong Evidence tab and a Low Confidence tab, so that all candidates can be viewed, regardless of whether they are called strong or not. The Strong Evidence calls should be considered the calls that are being made by the software. Low Confidence calls are primarily used for troubleshooting false negatives in the Strong Evidence category and are not to be considered positive calls. The information presented in the Strong Evidence and Low Confidence tabs shares the same format. There are a number of different criteria for the categorization of weak vs. strong, as described in subsequent sections.

In order to be called a strong fusion, the following criteria must be satisfied:

1. Minimum number of reads - In order for any candidate to be considered at all, there must be at least 5 breakpoint spanning reads that support the candidate. If there are not at least 5, it will not be further evaluated, and it will not be classified as either strong or weak (i.e. it will be absent from the results entirely). This cutoff can be adjusted by changing the MIN_READS_FOR_STRONG_FUSION parameter under "General Analysis Settings."
2. Presence of fusion in Quiver – If the fusion is found in the database of known fusions, called Quiver (http://quiver.archerdx.com), it will be called Strong (See *Appendix* for more details on fusion icons).

   The fusion is marked indicated with the bull's eye icon (◎) to indicate the fusion is in the Quiver database and the breakpoint is an exact match.

   The fusion is marked indicated with this icon (⛁) to indicate it is a known fusion but the breakpoint is not an exact match.

If a fusion is found in Quiver this overrides all other criteria and it will be reported in the Strong Evidence tab regardless of how weak the evidence might otherwise be, or what other criteria for a Strong fusion fail to be met. If the fusion is not found in Quiver, then the candidate can still be called Strong if it satisfies the rest of the criteria.

3. Percent GSP2 - Percent GSP2 is the proportion of breakpoint spanning reads that support the candidate relative to the total number of RNA reads spanning the breakpoint.

   Percent GSP2 needs to be at least 10% in order for the fusion to be considered Strong.

   Changing the STRUCTURAL_VARIATION_OF_GSP2_READS parameter under “Structural Variation Settings” adjusts this cutoff.

   If a candidate fails to meet the Percent GSP2 cutoff, it will be annotated with the (⚙) icon and be placed in the Weak category.

4. Minimum unique start sites - Within the population of breakpoint spanning reads that support the candidate, there will be a distribution of unique start sites.

   There must be at least 3 unique start sites to be considered Strong.

   Changing the MIN_UNIQUE_START_SITES_FOR_VALID_STRUCTURAL_VARIATION parameter under "Structural Variation Settings" adjusts this cutoff.

   If a candidate fails to meet the unique start site cutoff, it will be annotated with the (⚙) icon and placed in the Weak category.

5. Unless it is found in Quiver, a candidate must not trigger any of the conditions found in the Absence of Negative Evidence section to be called Strong Evidence. With one exception, if any of the negative evidence criteria are met, the candidate will be called Weak.

Negative Evidence Criteria

In order to be called a strong fusion, the following criteria must **not** be met (unless the fusion is present in Quiver):

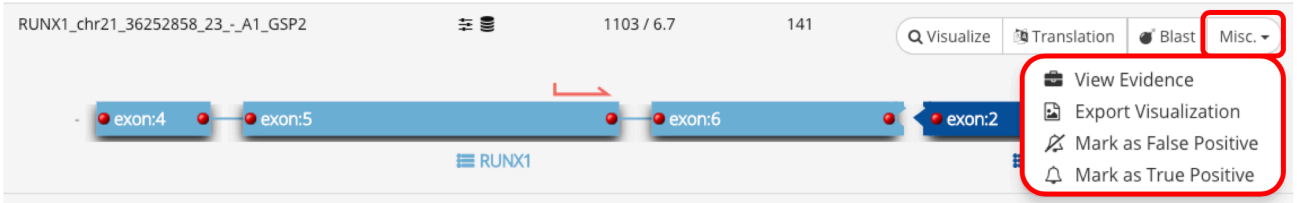
- Exon-Intron Fusion – If the fusion sequence on one side of the breakpoint is found to be entirely intronic (which is indicative of a DNA mispriming event), the fusion will be classified as exon-intronic.
  - This is to distinguish such events from those that utilize an intronic cryptic splice site resulting in just an internal portion of the fusion sequence corresponding to an intron, such as can be found in a common ALK-EML4 variant. Exon-intronic fusions are indicated with the (🔗) icon and placed in the Weak category.
- Mispriming – If significant sequence similarity is found between the fusion partners, the event is likely to be due to mispriming. Additionally, if the fusion breakpoint is less then 3bp from the GSP2, then similarity is assumed.
  - Likely off-target mispriming events are indicated with the (⚙) icon and placed in the Weak category.
- Paralogs – Archer Analysis compares the identities of the fusion partners with a list of known paralogs taken from the Ensembl database.
  - Known paralogs are indicated with the (⚙) icon and placed in the Weak category.

- Low Confidence – Annotation of each fusion consensus is performed by aligning the sequencing to the human genome with BLAST. The quality of these alignments, and the confidence of resulting annotation, depends on alignment length and repeat content of the sequence.
    - Events with a low confidence annotation are indicated with the ( ) icon and placed in the Weak category.
- Cross-Contamination – If a low expressing fusion candidate shows significant similarity to a high expressing fusion in the same analysis, it will be considered the result of intra-run cross-contamination.
    - Likely intra-run cross-contamination events are indicated with the ( ) icon and placed in the Weak category.
- Transcriptional Read-Through – Fusion transcripts of interest are generally derived from a genomic translocation event. However, fusion transcripts can also arise from failure of the cell to properly terminate transcription from a gene such that transcription continues on into the next gene downstream (if it is on the same strand).
    - Transcriptional read-through events are indicated with the ( ) icon and placed, by default in the **STRONG** category.
    - Transcriptional read-through events are placed in the **STRONG** category by default because by representing actual molecules produced in cells, they are technically not false positives.
    - Transcriptional read-through events can be made to appear in the **WEAK** category by changing the value from "OFF" to "ON of the parameter CALL_TRANSCRIPTIONAL_READTHROUGH_EVENTS_WEAK under the "Structural Variation Analysis Settings."
    - This is the only criterion in the "Negative Evidence" category that can be configured in this way.

Finally, if a fusion has been found to pass all the Positive Evidence criteria (except for the presence in Quiver, which is optional), and does not trigger any of the Negative Evidence criteria, it will be indicated with the ( ) icon.

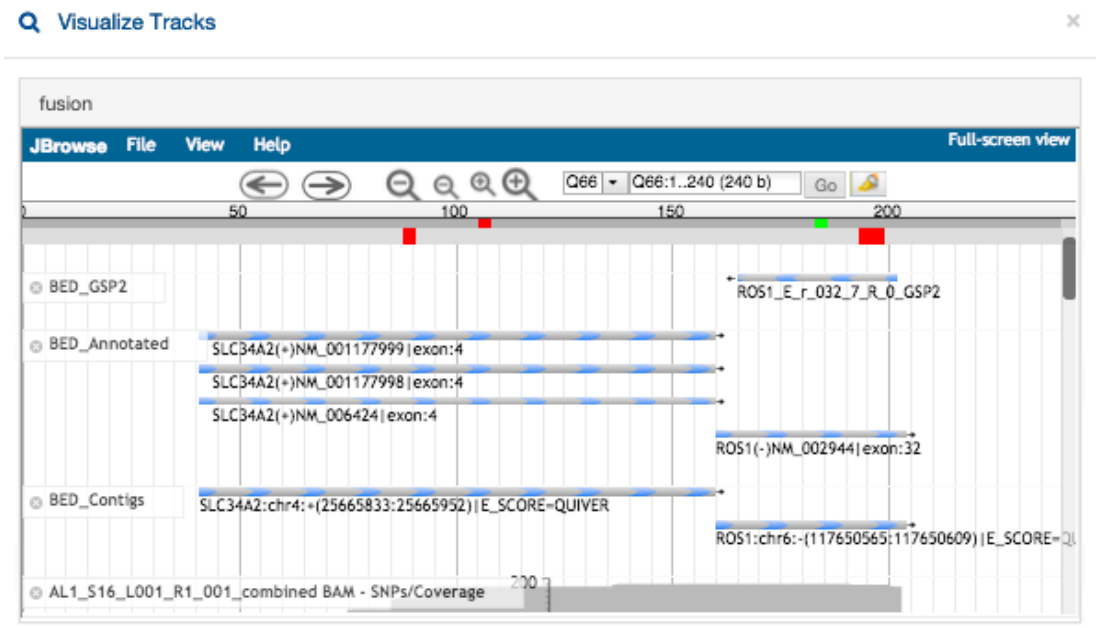
User-Annotated False and True Positives

It is possible to override the positive and negative evidence criteria by annotating specific fusions (according to breakpoint). Such annotation will be recorded and used in a manner similar to Quiver entries, and will trump all other criteria:

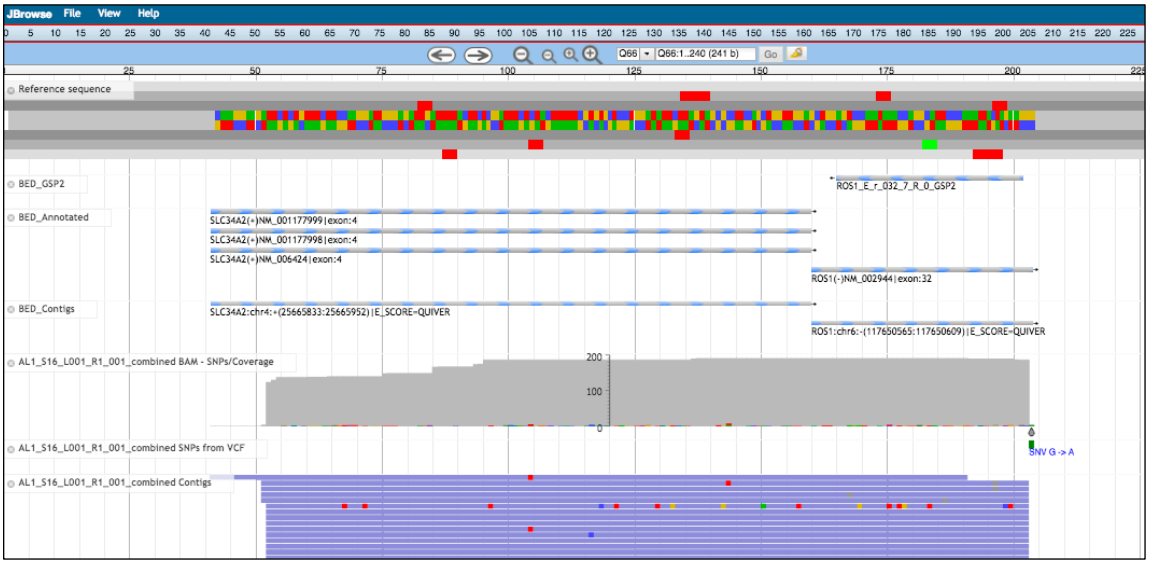


Visualization of Fusion Candidate-Supporting Reads in JBrowse

To verify the accuracy of the fusion calls, it is possible to visualize the reads supporting the breakpoint and fusion call (or wildtype call). Click the Visualize link next to the fusion of interest and a JBrowse window will open up with the supporting data.



The JBrowse viewer is provided in a small dialog box but clicking the "**Full-Screen View**" in the top right corner will open the genome browser in its own window or tab in the web browser.



The genome browser shows the constructed consensus sequence of the fusions as the reference sequence to which the de-duplicated reads supporting the fusion are mapped.

Tracks in the JBrowse View for Fusions

There are several tracks in the JBrowse view and a description of each is provided below.

BED_GSP2 – A track showing the location of the gene specific primer 2 (GSP2) used to detect the gene fusion.

BED_Annotated – A track showing the gene and exon number annotation for each of the fusion partners.

If a gene has multiple transcripts/isoforms, each of the possible isoform annotation is shown as a separate line.

BED_Contigs – The different contigs that make up this fusion**.** The annotation of each of the regions in this track contains the name, the location on the hg19 genome as well as the BLAST E score on which the gene annotation was based.

Fusions are not aligned against hg19 directly. First, a fusion reference is created from the hg19 annotations. This is done in order to see reads aligned across the breakpoint.

***_coverage –** The coverage track for the reads supporting the fusion

*** Contigs –** The unique (de-duplicated) reads supporting the fusion

SNPs from VCF – Any differences between the consensus sequence and the reference are listed here as a VCF track. In some cases there can be a large insertion in the alignments that typically corresponds to a segment of the fusion contig that could not be annotated to the genomic reference.

Protein Translation prediction

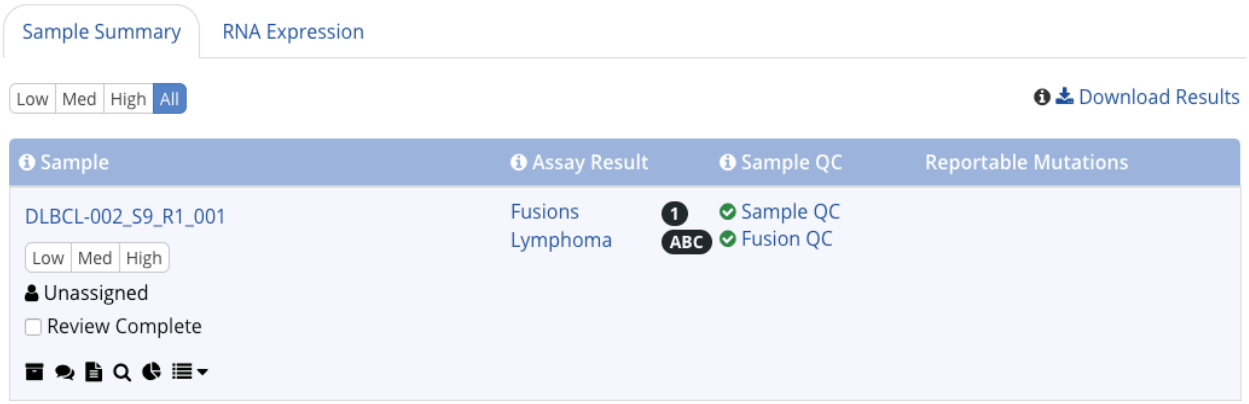
For gene fusions, the Archer Analysis software determines if the resulting protein is in or out of frame using all combinations of the fusions candidate isoforms and creates an amino acid sequence that can then be subsequently BLAST'ed against the human protein database using the web interface to the NCBI BLAST application, to provide more information about the fusion protein.

Clicking the Translation box provides a new window with the prediction results and a link to the BLAST application.

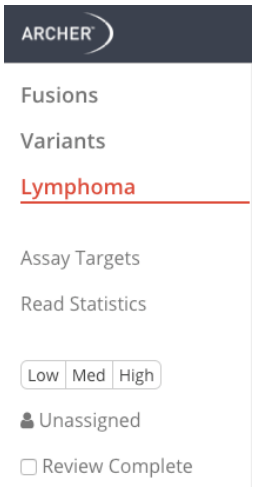
Lymphoma Cell of Origin (COO) Pipeline and Results (Experimental)

This pipeline is designed to identify the cell of origin subtype from a diffuse large B-cell lymphoma (DLBCL) sample. It is intended to run on known lymphoma samples, not determine if an unknown sample is a lymphoma sample. The classification is based on the work in “Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue” (PMC3931191). The analysis pipeline will occur after the standard primary analysis and is based on calculated gene expression. The pipeline is automatically enabled for panels designed to detect lymphoma COO.

The results of the pipeline displayed in the “**Assay Results**” column of samples is located on the “**Sample Summary**” page and will provide the Lymphoma results: ABC, GCB or Unclassified. Select the link in the “**Assay Results**” column for more detailed information on the results.



By selecting "**Lymphoma**", this will take you to the lymphoma "**Overview**" page that displays the subtype metric, subtype probability and a lymphoma gene expression heat map. You can also navigate to the Lymphoma overview page by selecting "**Lymphoma**" on the left side of the sample details page.



Lymphoma COO Metrics:

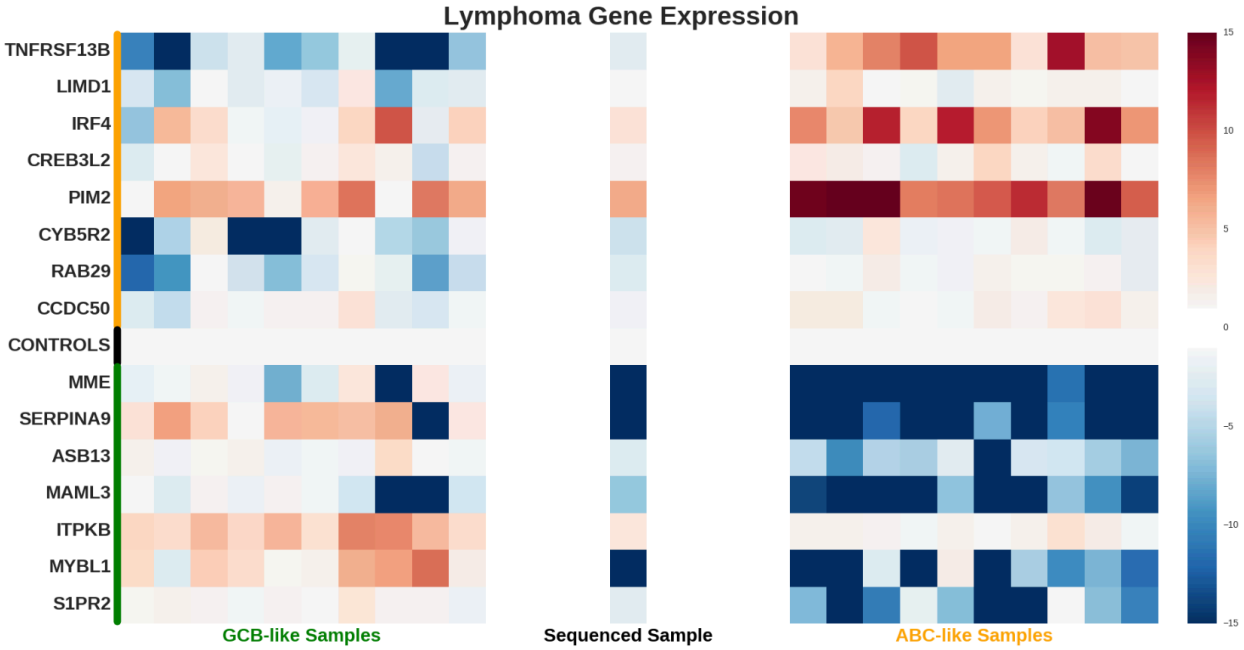
- **Subtype Metric** – The reported lymphoma cell-of-origin subtype (GCB, ABC or Unclassified).
- **Subtype Probability** – Estimated probability of subtype classification.
- **Lymphoma Gene Expression Heat map** – This is a heat map of the gene expression considered in the COO algorithm. The results of the sequenced sample are plotted against known ABC and GCB reference samples.

Overview

Subtype Metric	Subtype Probability
GCB	0.925

Export Data (tsv)

Heatmap



SNP/InDel (RNA and DNA)

The DNA small variation workflow detects SNPs (single nucleotide polymorphisms) and InDels (insertions and deletions) in your data. This can be utilized in addition to other workflows. The software will produce a variant summary with customizable filters that can be used to filter variants of interest. These filters can be adjusted in the interactive table and saved for future analyses. The user may select which attributes of the variants to show using basic logic gates provided in the software. When satisfied with the results, the data may be exported into a PDF report or downloaded as a TSV (tab separated values).

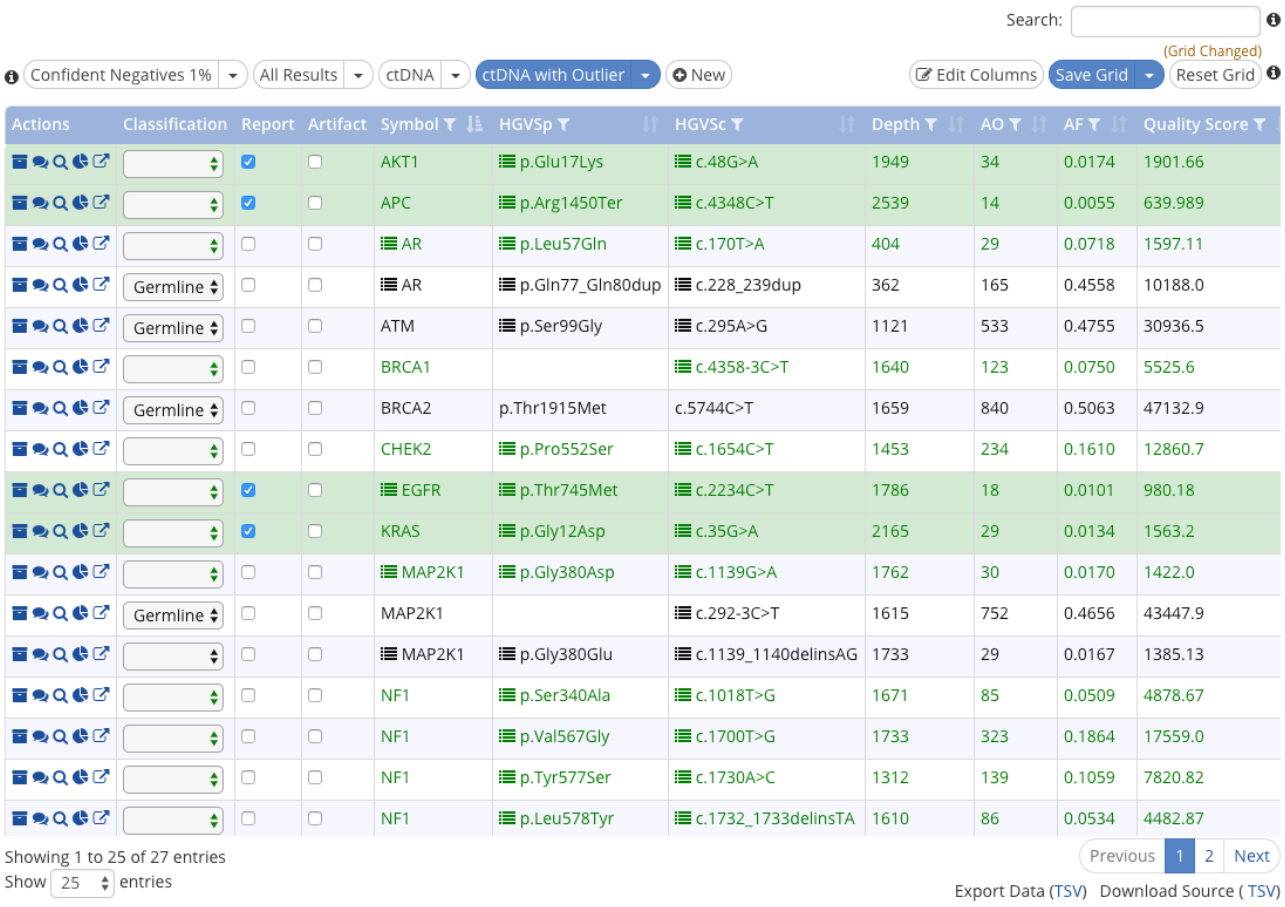
Targeted vs. Non-Targeted Calling

In the "**Detailed Summary**" section “Variant Summary”, each of the targeted mutations will be listed – not just those that were found to be actually present in the sample – with their status listed as “NA”.

The Variant Summary page for a sample that used the Targeted Mutation option shows only those variants that were defined in the target variant call format (VCF) file and if “Include Non-targeted” is turned OFF. The names of the variants consist of the gene name and amino acid mutation as defined in the target VCF file.

General Information About the Variant Summary Grid

The Variant Summary tab in the **Sample Details** page will show the results of the variant detection as shown below:



The "**Variant Name**" column will be empty for variants that were not detected using targeted mutation analysis.

In addition to the variant calls and the basic statistics about the coverage etc. the Archer Analysis software also provides more detailed information about the potential effect of the variant. This is achieved using the Variant Effect Predictor tool, developed by the EBI at ENSEMBL (http://www.ensembl.org/info/docs/tools/vep).

Editing and Saving Variant Grid Columns (across one or more samples within a job)

Users have the ability to edit the Variant Grid by visualizing the column headers as they see fit. Once the Variant Grid columns are rearranged, those settings can be saved across a single sample or multiple samples with a job.

Filtering Results

Each of the columns form the Variant Summary can be filtered either by selecting a value from the drop down menu or by typing in the search commands in the text box.

Columns that use a drop down menu have arrows; selecting the dropdown shows the values that are found in the columns.

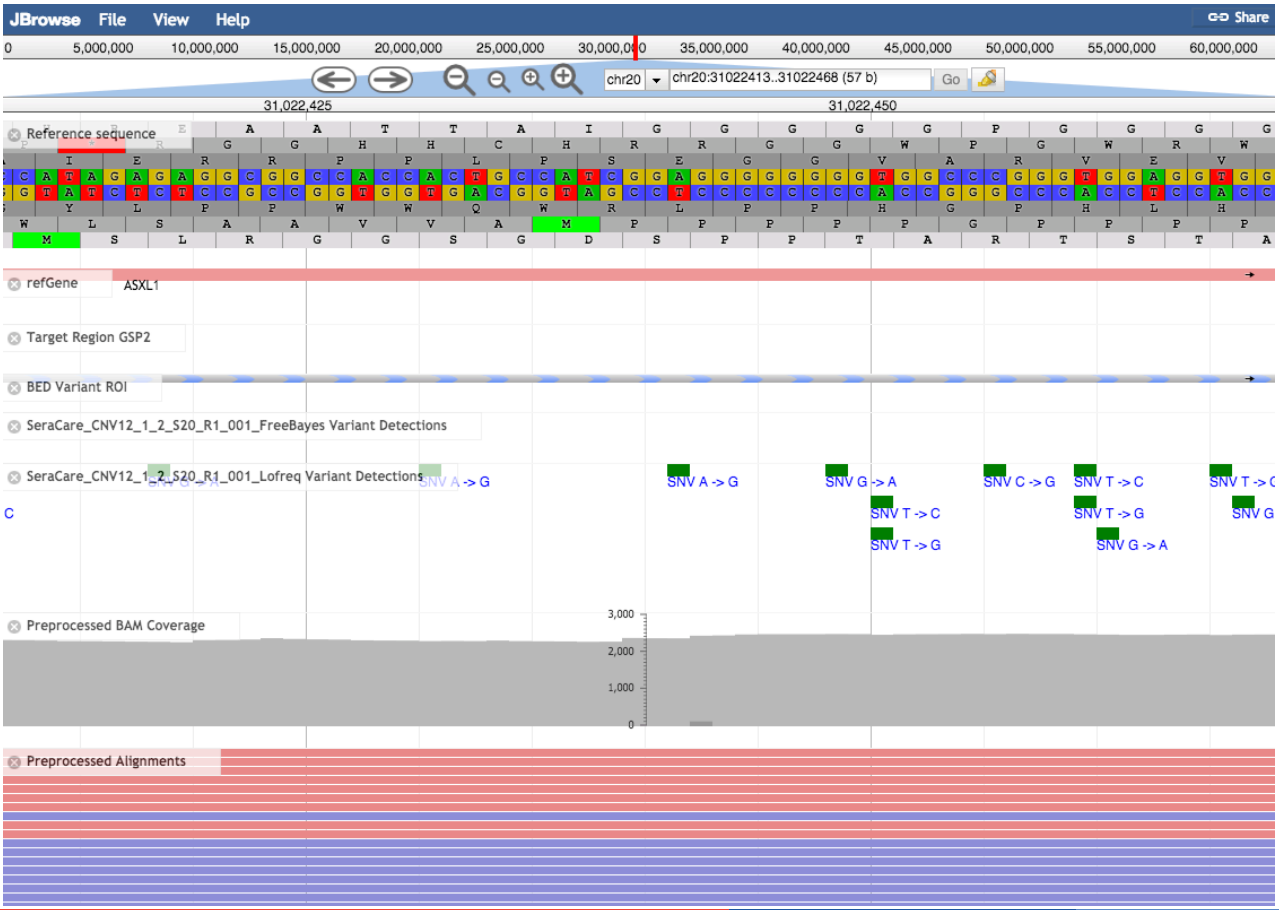
The columns with text boxes can be used to enter search text and any rows not containing the search text will be removed from view.

The search text can contain special operators to make more complex search queries such as combining fields with Boolean statements or, for columns containing numerical values, the rows that contain a value that is less than or more than some search value can be selected as well.

The information currently displayed in the Variant Summary can be downloaded by selecting the "**Export Data (TSV)**", or the unfiltered data is available by selecting "**Download Source (TSV)**", below the variant summary page.

Visualization of Variant-Supporting Reads Using JBrowse

The reads supporting a variant can be directly visualized from within the Archer Analysis software using JBrowse, by clicking the location link in the Genomic Location column (e.g. chr2:1481231). This brings up a separate window with JBrowse, focused on the selected location.



There are a number of tracks in the JBrowse window:

refGene – A track showing the genes and each of their transcripts from the RefSeq database

Right-click on the transcript to see it's name and follow the link for a search of the transcript at the NCBI website.

Cosmic – A track showing the location and identifiers from the COSMIC mutation database. Right-click on the feature to follow the link of the identifier at the Sanger website.

SNPs from VCF – A track showing the variants detected (or targeted)

Merged BAM Coverage - Both Read 1 and Read 2 are combined into a single BAM file.

Preprocessed BAM Coverage – BAM coverage after filtering and InDel realignment.

Copy Number Variation (DNA Only)

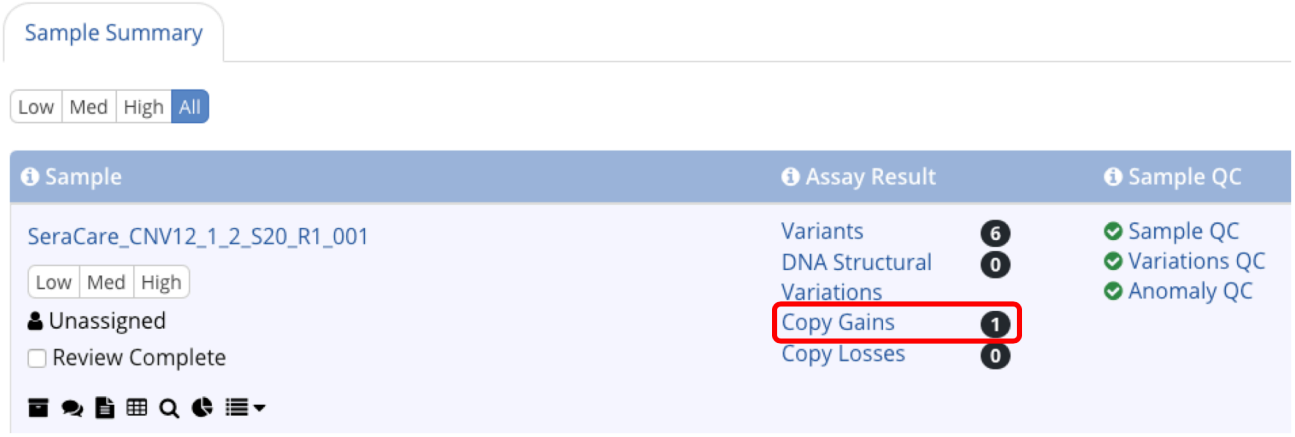
CNV detection can be selected in addition to other assay types if the user has a kit that is designed for CNV detection. This assay uses one or more control samples, if present, to compare copy number in normal versus diseased samples. If no controls are present, all samples are used as a baseline. Selection of the proper normal sample(s) is described in *Selecting Normalization Parameters for CNV* section. The analysis software uses this information to produce both a readout and a visualization of copy number, along with a p-value to convey confidence in the calls.

If no matched case-control samples are available, the use of multiple control samples could increase the sensitivity. The group of control samples will inform the algorithm about the natural variability of the coverage and GSP behavior. This could result in more accurate CNV calling. Only the samples that are marked as tumor samples will show CNV results. Samples marked as normal will not display data in the "**Assay Results**" column.

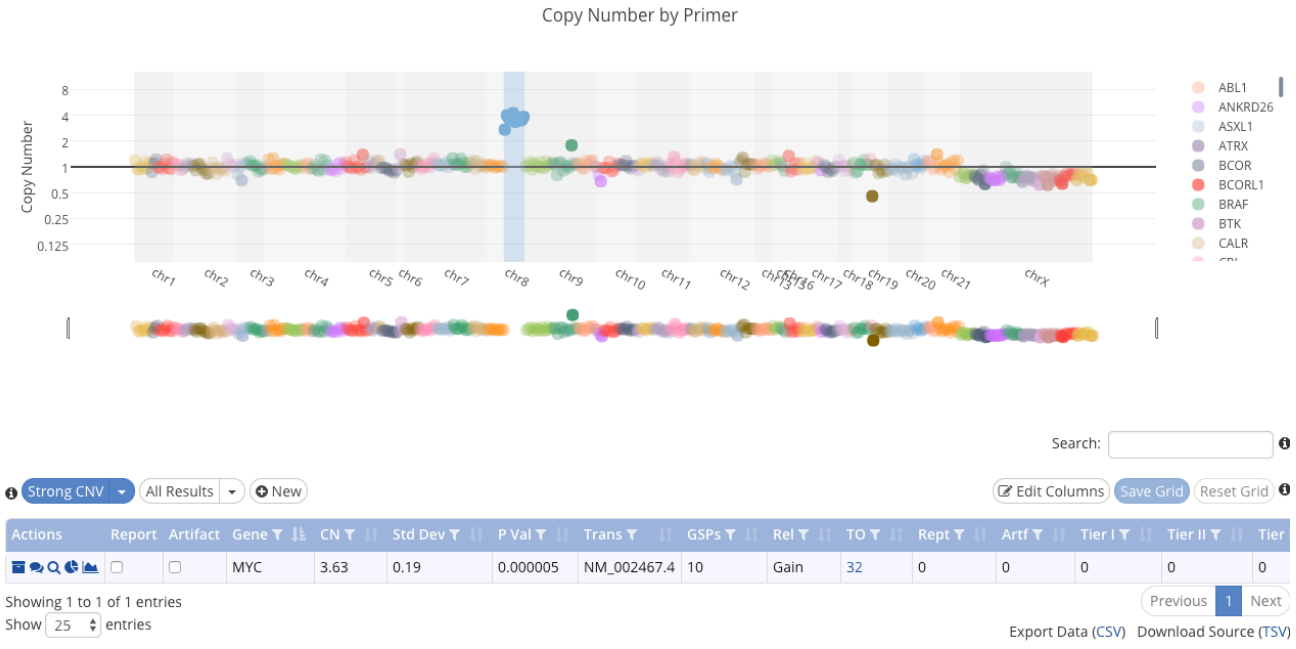
Data Summary

For DNA Copy Number Variation analyses, the **Job Detail** page will list the results of the CNV analysis by highlighting the genes that show significant copy gains or losses:

Report partial CNV where at least three adjacent primers show significant copy gains or losses.

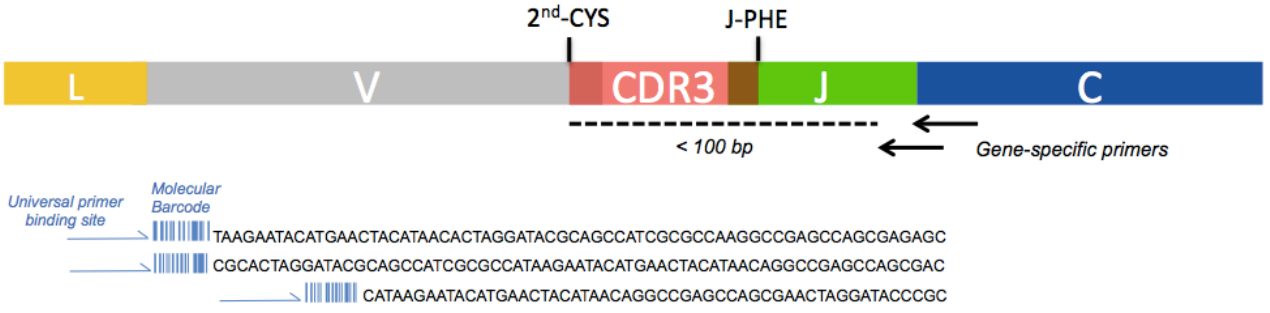


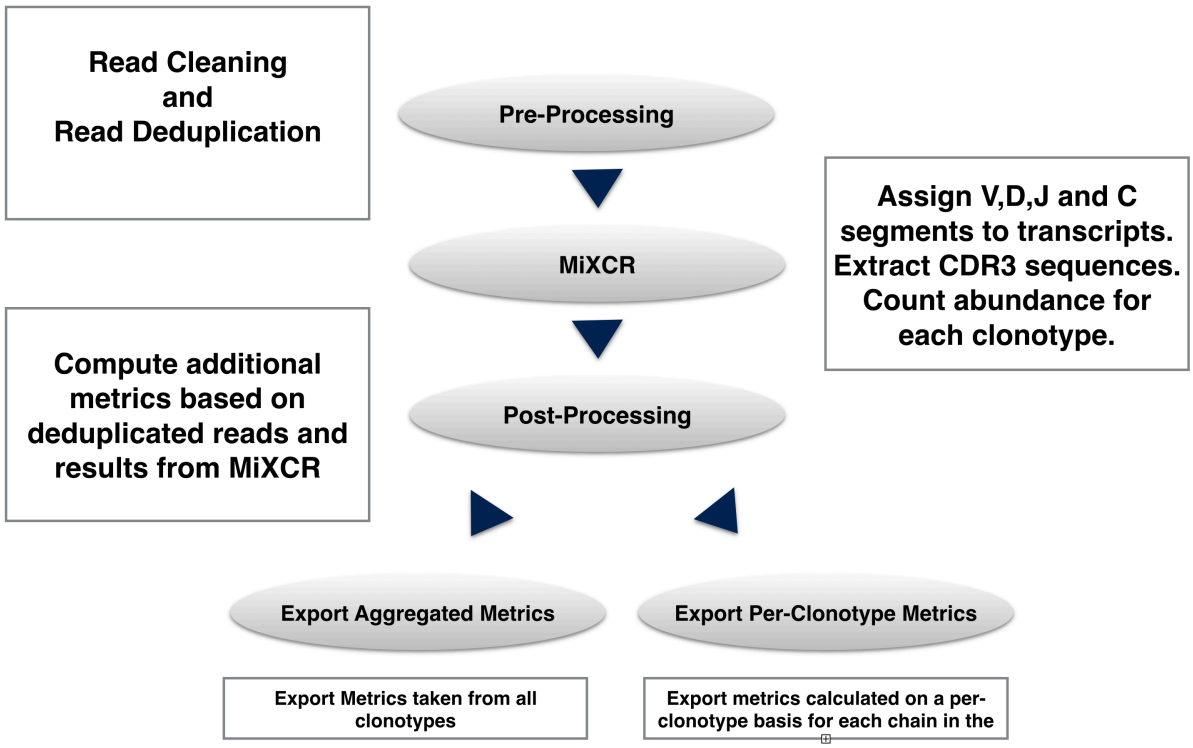
And then further detail of the CNV data on the **Sample Detail** page:



Immune Repertoire (RNA Only)

The Immune Repertoire pipeline provides an overview of the immune receptor population expressed by IG and TCR clonotypes within a sample. Reads are pre-processed (i.e. adaptor trimming, deduplication, error correction), and then profiled using a third party tool (MiXCR) with some post-processing. Please note, error correction is automatically enabled for the pre-processing of this pipeline. MiXCR processes Archer Immunoverse data to assign V, D, J and C regions to the unique fragments, as well as extracts the primary structure of the CDR3 sequence. It then clusters fragments with similar CDR3 sequences together thus providing a consensus sequence for the CDR3 as well as an abundance count of number unique fragments supporting the CDR3.

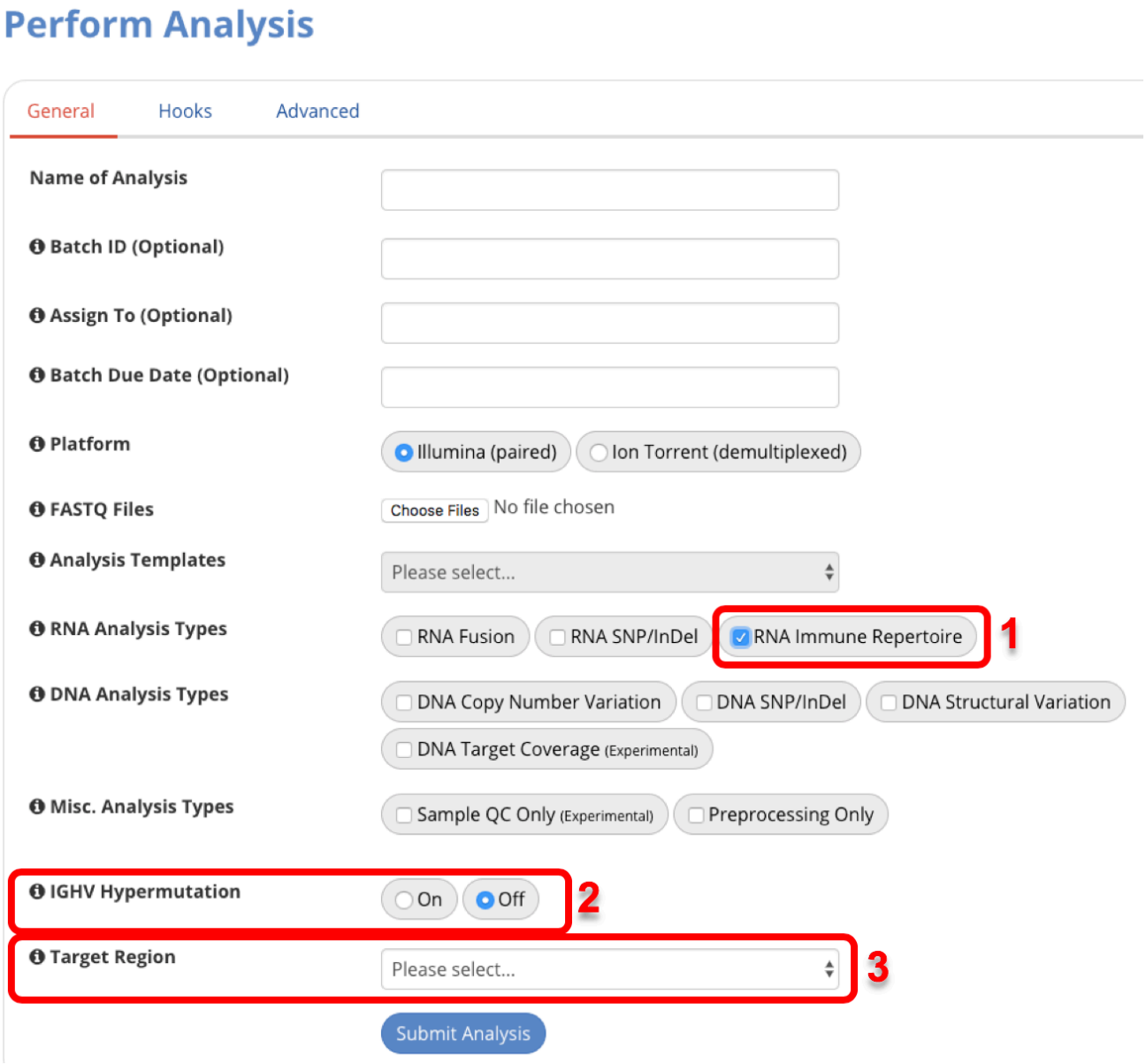




Job Setup for Immune Repertoire

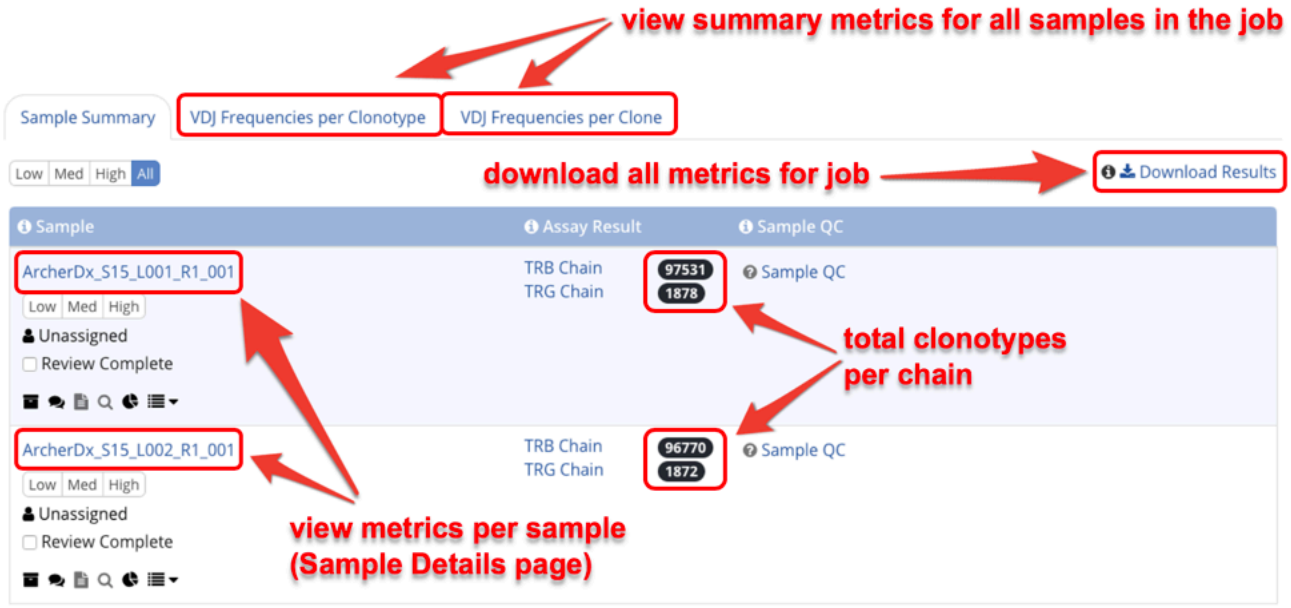
Start the analysis of your samples via the "**Perform Analysis**" page (see *Job Setup and Management for general steps*). Elements of this process that are unique to this pipeline are:

1. Select Analysis Type = "RNA Immune Repertoire"
2. Choose whether to calculate V-segment Hypermutation status (for top 5 most frequent clonotypes in IGH libraries)
3. Select the appropriate Target Region file (GTF file) for the specific Immunoverse panel used to generate your library (see Job Setup and Management for further details)



Job-level Results

After the job is processed, results can be viewed and downloaded via the **Job Summary** page. The GUI reports many, but not all, of the MiXCR and post-processed metrics, on both per job and per sample bases. Comprehensive results, in tab-delimited format, can be obtained by utilizing the "Download Results" button on the Job Summary page, as per other pipelines in Archer Analysis:



Sample-level Results

Access the **Sample Details** page by clicking on sample name on the **Job Summary** page. There are two tabs on this page:

- "**Read Statistics**" overviews key metrics before and after processing read data
- "**Clonotypes**" provides an in depth look at clones and clonotypes

Molecular Barcode Statistics

Total Fragments	Fragments with Complete Adapter	Number of Reads After Trimming Adapters
3,500,000	3,418,281	2,702,650

Export Data (tsv)

Read Statistics

Type	Total Fragments (# / %)	Mapped (# / %)	Pass Alignment Filter (%)
All Fragments	2,702,650 / 100.0	2,527,198 / 93.5	93.5
Unique Fragments	570,273 / 21.1	532,596 / 93.4	93.4

Export Data (tsv)

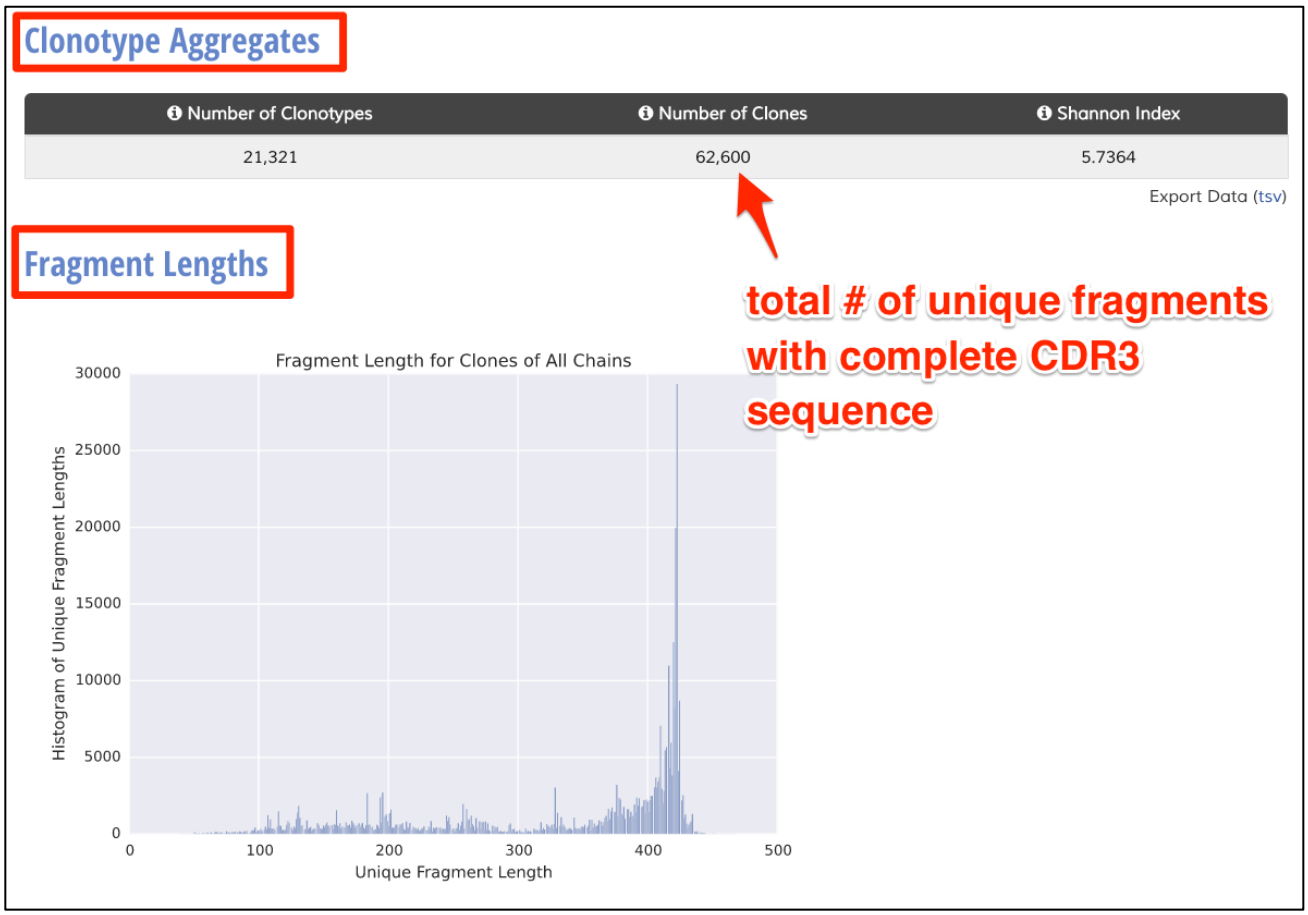
DNA/RNA Fragment Lengths

DNA Median Fragment Length	DNA Mean Fragment Length	RNA Median Fragment Length	RNA Mean Fragment Length
145.0	164.4	213.0	204.1

Export Data (tsv)

Sample Details page → Read Statistics tab

The top sections of the “Read Statistics” tab provide metrics similar to other Analysis pipelines. Unique information for this pipeline is provided in two sections at the bottom of this tab (“Clonotype Aggregates” and “Fragment Lengths”):



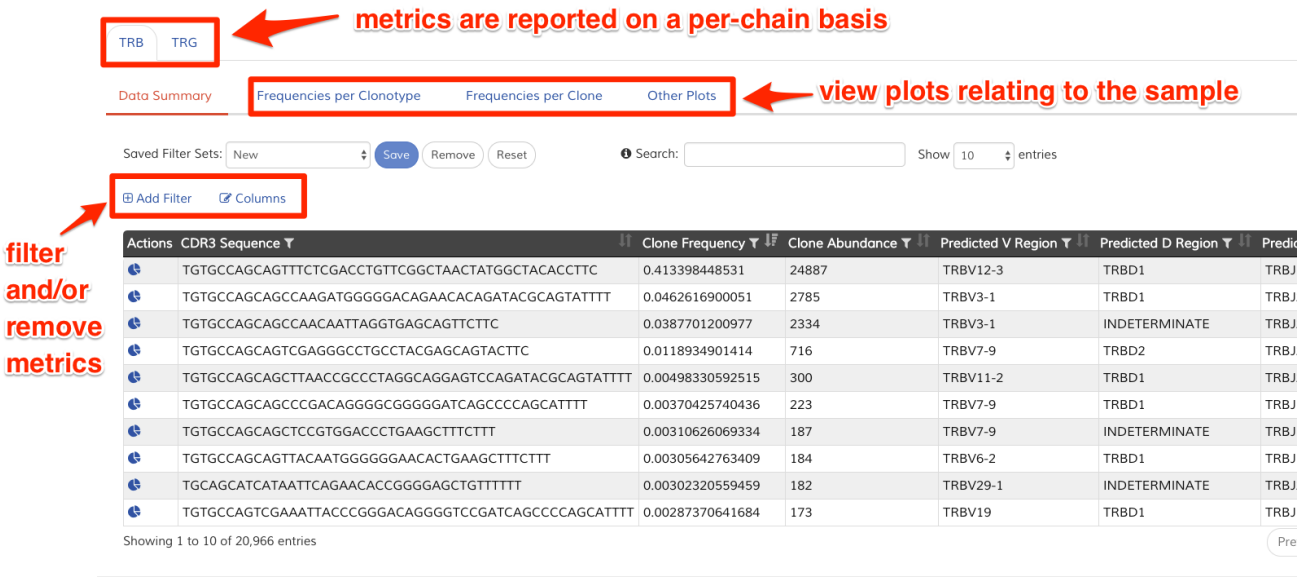
Sample Details page → Clonotype Summary tab

Clicking on the metrics under the "**Clonotypes**" Column brings up a new page with new tabs for each chain included in the Archer panel relevant to the current sample. Each of these tabs provides a grid to investigate all clonotypes (CDR3 sequences) found in the sample.

The clonotype grid functions in a manner analogous the variant grid of the **SNP/InDel** pipeline of Archer Analysis; each column represents a metric that can be filtered/sorted against, and visibility toggled on/off. Additionally, data from the grid can be exported as a tab-delimited file, by clicking links at the bottom right-hand corner:

- "**Export Data**" exports only what is currently displayed in the grid
- "**Download Source**" downloads all clonotype data (i.e. ignoring column filtering and visibility).

In addition to the grid, plots (histograms and heat maps) of clonotypes, associated clones, as well as V, D, and J segments, can be accessed on this page as well:



Per Sample Plots

Below details the per-sample plots that can be accessed via the "Clonotypes" section:

In the "Clonotypes" section, there are four tabs:

1) Data Summary: Contains the grid with metrics available on a per-clonotype basis for the selected chain.
2) Frequencies per Clonotype: Contains visual displays of segment usage frequencies of each clonotype:
   a) Predicted V-Region Frequency: frequency among all V-regions found per-clonotype.
   b) Predicted D-Region Frequency: frequency among all D-regions found per-clonotype.
   c) Predicted J-Region Frequency: frequency among all J-regions found per-clonotype.
   d) Predicted C-Region Frequency: frequency among all C-regions found per-clonotype.
   e) CDR3 Translation Frequency: Frequency of CDR3 translations amongst all clonotypes of the specified chain
   f) Productive CDR3 Translation Frequency: Frequency of CDR3 translations amongst all clonotypes of the specified chain of only productive CDR3 translations.
   g) CDR3 Sequence Frequency: Frequency of CDR3 sequences amongst all clonotypes of the specified chain
3) Frequencies per Clone: Contains visual displays of segment usage frequencies of each clonotype the plots that can be selected are:
   a) Predicted V-Region Frequency: frequency among all V-regions found per-clone.
   b) Predicted D-Region Frequency: frequency among all D-regions found per-clone.
   c) Predicted J-Region Frequency: frequency among all J-regions found per-clone.
   d) Predicted C-Region Frequency: frequency among all C-regions found per-clone.
   e) CDR3 Translation Frequency: Frequency of CDR3 translations amongst all clones of the specified chain
   f) Productive CDR3 Translation Frequency: Frequency of CDR3 translations amongst all clones of the specified chain of only productive CDR3 translations.

g) CDR3 Sequence Frequency: Frequency of CDR3 sequences amongst all clones of the specified chain

4) Other Plots:
   a) CDR Length: A sequence length histogram of all unique CDR3 sections found in the chain
   b) Raw Fragment: A fragment length histogram of CDR3 positive reads.
   c) Raw V segment: Histogram of the number of base pairs covered in the V segment per read
   d) Unique Fragment: A fragment length histogram of CDR3 positive deduplicated reads.
   e) Unique V segment: Histogram of the number of base pairs covered in the V segment per deduplicated read.
   f) V & J usage: A heatmap detailing co-occurrences of V and J usage per clonotype.
   g) V usage: A histogram of V segment usage of the 100 most abundant clonotypes
   h) D usage: A histogram of D segment usage of the 100 most abundant clonotypes
   i) J usage: A histogram of J segment usage of the 100 most abundant clonotypes
   j) C usage: A histogram of C segment usage of the 100 most abundant clonotypes

Post-processing after Archer Analysis

All data presented in the Archer Analysis GUI, as well as all metrics generated by MiXCR, can be exported for further analysis (e.g., via R, Excel, etc.). This data can be accessed by clicking the "User Defined File Download List" link on the **Job Summary** page. This will initiate download of a zip file to your local machine, named after the job number. After extracting the archive, data for each sample can be found in its own subfolder, named as (with "*" as a placeholder for the sample name):

***.molbar.trimmed.deduped.Immune_Repertoire_Results**

Each sample subfolder contains all plots (as PDFs) and raw text output (as tab-delimited files). Further details on this output can be found in Immune Repertoire Files.

Outlier Detection

Outlier detection is an algorithm that identifies variants with allele fractions that are statistically significant in a sample compared to a set of samples. If Normal Data Sets are not used (see section), Outlier detection will first identify positions in samples with unusually high allele fractions and reserve them from noise profiling on the hypothesis that these data are true positives. The algorithm then models the noise on a per-position basis across samples and then compares specific sample variant calls to the noise profile to generate statistics to discern true variants from noise. This feature helps provide greater confidence in variant calling and increased sensitivity at low allele fractions while increasing specificity, providing confidence in variant calling.

When Outlier Detection is enabled, the feature can be used in two different ways: intra-job and Normal Data Sets. Intra-job models the noise at each sequenced position across all samples within a job. Normal Data Sets allows a user to designate normal samples so that noise is modeled only using these samples.

When designating a Normal Data Set of samples, it is critical to ensure that those samples do NOT contain the variations you wish to measure in the assay. If they are present, the ability to accurately estimate the background noise of those variations will be diminished - subsequently lowering the sensitivity of those variations in the unknown samples. It is recommended to use multiple samples

from multiple sources, so that variants or SNPs specific to individual samples do not skew the noise modeling.

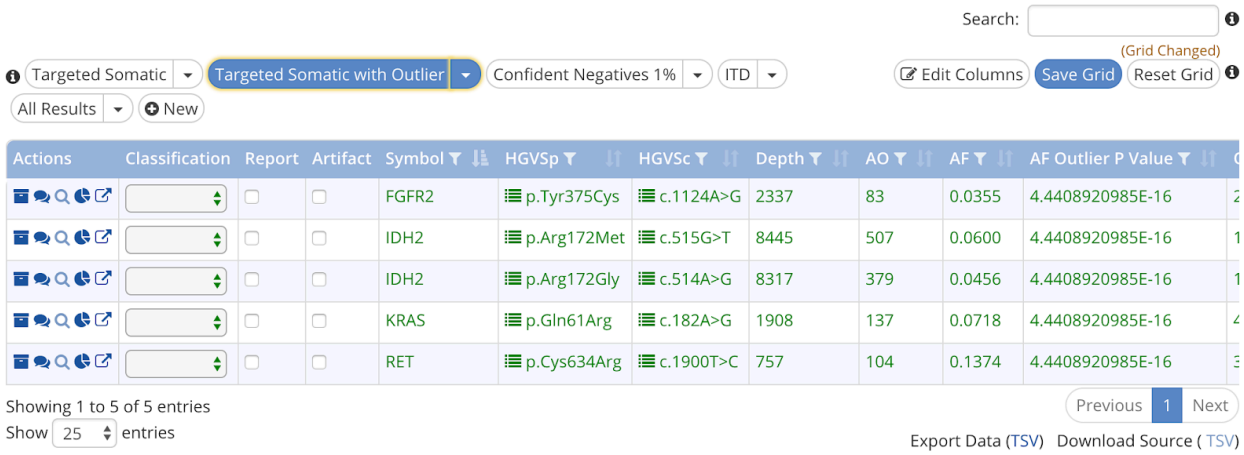
Enabling/Disabling Outlier Detection

From the menu bar, select the "**User Email**">"**Settings**" to access the "**User Settings**". Scroll down to the "**Variations/Mutations Settings**" where the toggle for "**OUTLIER_DETECTION**" can be found.

Outlier Detection is enabled by default, but can be turned off if the user wants to disable the feature. If you make a change to the settings, you will want to click "**Update Settings**" located at the bottom of the "**User Settings**" page.

Using Outlier Detection in the Variant Grid

For variant calling, users have the ability to filter variants using Outlier Detection p-value metric. By default, this filter set is the same as the original variant filter set except the software adds the "**AF Outlier P Value**" with a default p-value. By selecting the filter set with Outlier, only variants with significant p-values will be reported in the Variant Grid.



Note, Outlier Detection will not be beneficial to run on Germline panels. To save analysis runtime, manually disable the feature in "**User Settings**".

The "**Confident Negatives**" filer set is useful to determine which genomic positions can be reported as reference. By default, this filter set displays positions with 95MDAF below 1% that also have insignificant AF Outlier p-values.

Key Outlier Detection Metrics

Intra-Job AF Outlier P Value (Intra-Job Allele Fraction Outlier P-value) - The probability this mutation was due to background noise (BN), as estimated across all samples in the same job. A low p-value indicates the AF of the called variant is significantly outside the BN, providing confidence the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Minimal Detectable Allele Fraction 95% (95MDAF) - The allele fraction (AF) at which a variant can be called with high statistical power. If the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called. This value will be the 95MDAF given the background of the normal data set if one was provided, otherwise it will be the 95MDAF of the intra-job comparison if outlier detection is on.

Normal Data Sets (SNP/InDel & CNV)

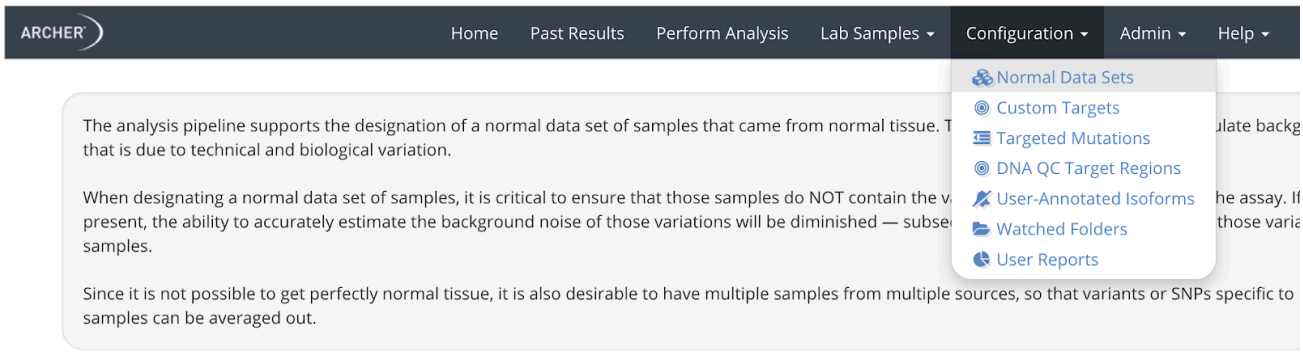
This feature supports the designation of a Normal Data Set of samples originating from normal tissue. These samples will be used to calculate background noise that is due to sequencing and enzymatic error.

When designating a Normal Data Set of samples, it is critical to ensure that those samples do NOT contain the variants you wish to measure in the assay. If they are present, the ability to accurately estimate the background noise of those variations will be diminished — subsequently lowering the sensitivity of those variations in unknown samples.

Since it is not possible to get perfectly normal tissue, it is also desirable to have multiple samples from multiple sources, so that variants or SNPs specific to individual samples can be averaged out.

For SNP/InDel Normal Data Sets, it is also important to ensure that the total depth of coverage among all samples is sufficient. For an example if a genomic coordinate only has 50x total coverage calculated among all samples in the normal data set, outlier detection will not be sufficiently powered for variants with an allele fraction of one in a hundred.

To access the Normal Data Sets configuration page, select the "**Configuration**" **>** "**Normal Data Sets**" menu item.



Normal Data Sets Terminology

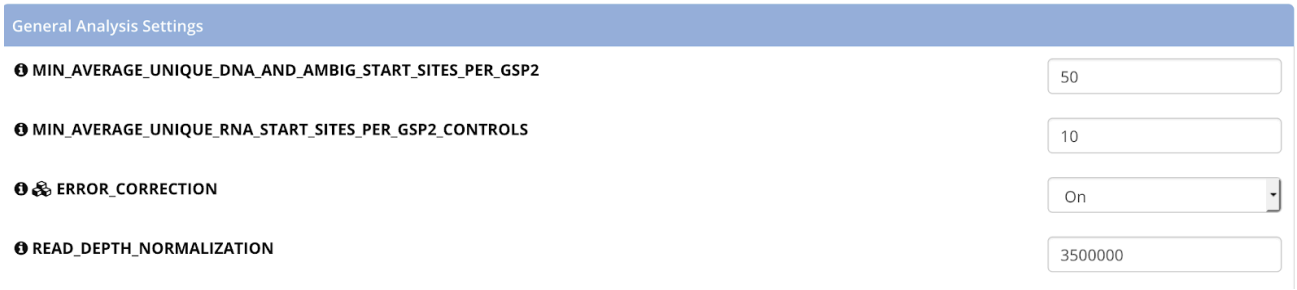
Committed Normal Data Set - An uncommitted Normal Data Set is one that can have samples added and removed. Being committed means that a job was run using the Normal Data Set. Once this happens the Normal Data Set is no longer allowed to have samples added or removed.

Deprecated Normal Data Set - A Normal Data Set that used to work on previous versions of analysis but because of a system upgrade the samples that are part of the Normal Data Set are now incompatible with the current state of analysis. Because of this the Normal Data Set cannot be used until it has been upgraded.

Performing Analysis to Create Normal Data Sets

A Normal Data Set is based on mutation type, sequencing platform, targeted region file and user settings. Therefore all samples that a user wants to be included in Normal Data Set must share these parameters.

The settings that must be the same are defined in the next section. To help users remember what these settings are, they are marked in the "**User Settings**" page with a Normal Data Set icon.



User Settings for Normal Data Sets

In order to add samples to a Normal Data Set, these "**User Settings**" must be the same across the samples:

- "ERROR_CORRECTION"
- "BARCODE_HAMMING_DISTANCE"
- "SKIP_DEDUPLICATION"

For a SNP/InDel Variant Normal Data Set in particular, these additional "**User Settings**" need to be the same as well:

- "MAPQ_THRESHOLD_FOR_VARIANT_CALL"
- "MIN_BASEQUAL_FOR_VARIANT_CALL"
- "DEEP_SHALLOW_THRESHOLD"
- "VARIANT_DOWNSTREAM_ROI_SIZE"

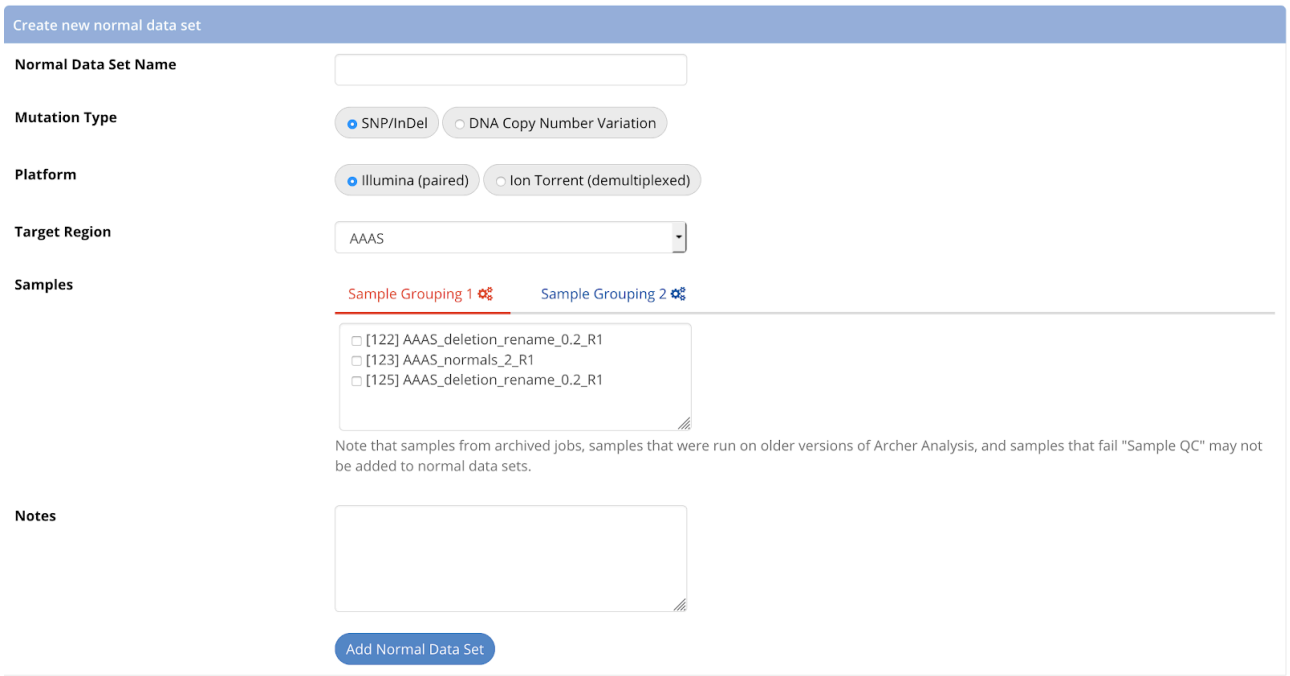
Creating a Normal Data Set

To create a Normal Data Set, select the "**Configuration**" **>** "**Normal Data Sets**" **>** "**Create Normal Data Set**".

When creating a Normal Data Set, users must provide:

- Normal Data Set Name
- Mutation Type (SNP/InDel or DNA CNV)
    - If samples were initially run with both SNP/InDel and DNA CNV analyses, and the user wants to create a Normal Data Set for both analysis types, the user must create two separate Normal Data Sets.
- Platform
- Target Region
- Samples
    - Grouping of samples per user settings, see below for sample visibility information
- Notes (optional)

Sample Visibility



Samples will be shown for selection once the target region is selected. In order for a particular sample to appear as an option the following requirements must be met:

- The platform and target region must match that of the job that contains the sample.
- The sample must not have errored out during processing.
- Overall sample QC must have passed for the sample.
- The sample must contain all the required files to be used within the current version of Normal Data Sets. As long as the job was run on the current version of the code this should never be an issue.
    - Any job that contains a sample(s) associated with a Normal Data Set must not be archived.

Samples will be grouped together based on the values of the concurrent settings listed above in the "User Settings for Normal Data Sets" section. In the screenshot above there are 2 setting groups listed. Hovering over the setting icon will show the values and how they differ from the default. Samples can only be selected from within a single group.

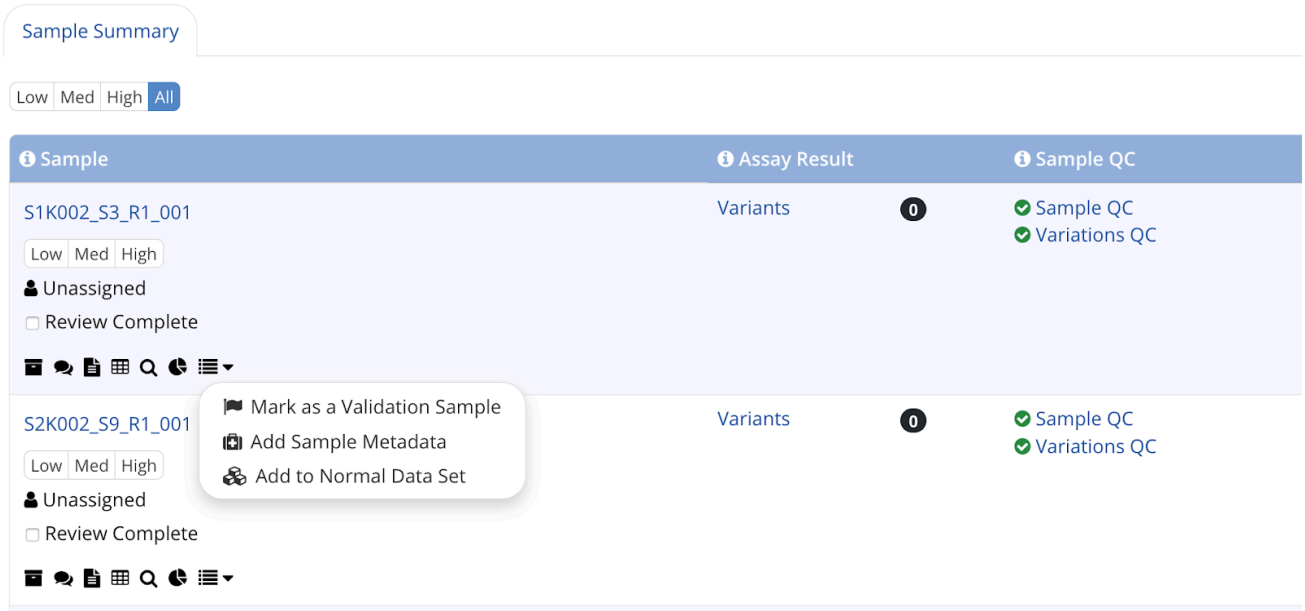
The following section explains an alternative way to add samples to a Normal Data Set which is useful to determine why a particular sample cannot be added to a Normal Data Set.

Adding a Sample to an Normal Data Set from Job Details

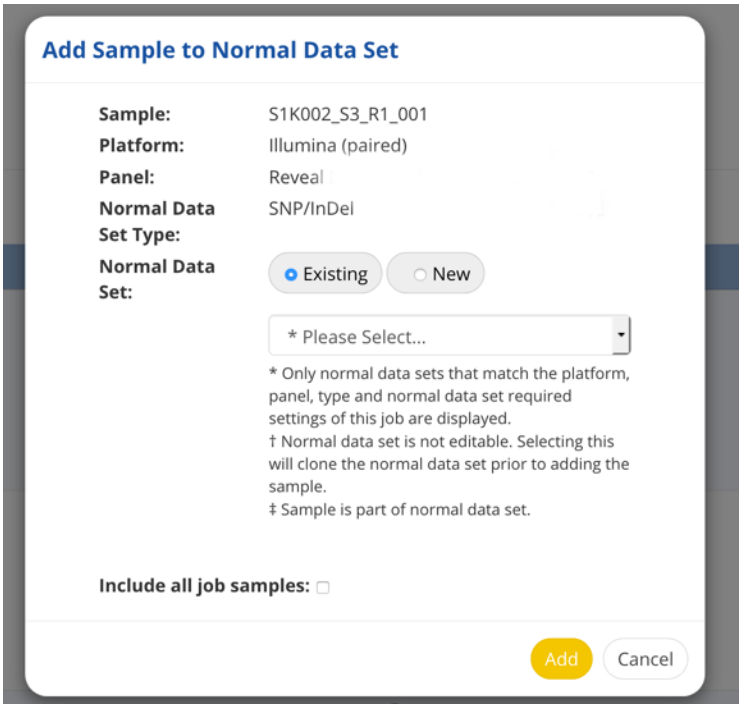
A user can add a sample to a Normal Data Set from the job detail page. This can be useful for the following reasons:

- This method will display the exact reason why a sample could not be added to a Normal Data Set.
- If you know the job that the sample is part of, it may be easier to find the sample rather then sifting through all sample IDs and names in the selection scroll box.
- All samples within a particular job can be added with a single checkbox.
- Streamlines cloning and adding a sample.

This method is accessed via the job details page where each sample is listed with its overall results. Click on the drop down item list icon.



Then click 'Add to Normal Data Set':



If this sample cannot be added to any Normal Data Set the reason will be displayed in place of the existing Normal Data Set drop down. Also listed is the platform and panel for the job; only Normal Data Sets defined with those values will be shown in the existing Normal Data Set drop down. In addition, only Normal Data Sets whose settings match those of this job's will be visible.

The first option is to add the sample to an existing Normal Data Set. This option gives a drop box allowing the user to select an applicable Normal Data Set. There might be an icon preceding the sample name; this is explained in the legend below:

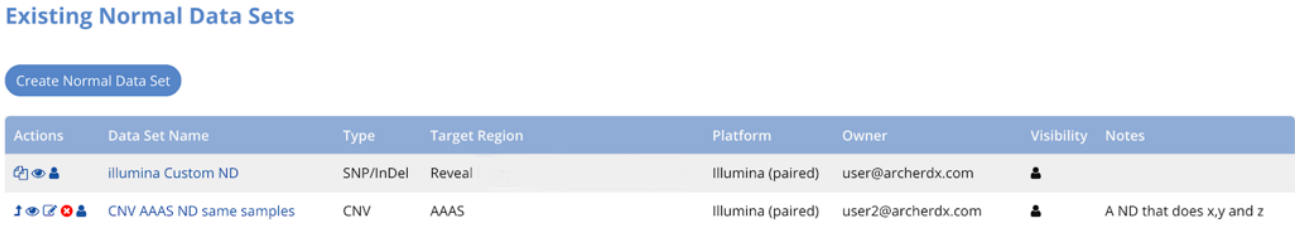
- No icon. The Normal Data Set is editable. Selecting one of these Normal Data Sets will add the sample to the Normal Data Set and redirect the user back to the job details page giving a *successful* message.
- † Normal Data Set is not editable because it has already been committed. Selecting one of these will clone the Normal Data Set prior to adding the sample and the user will be directed to the "**Create Normal Data Set**" page providing them the option to change the name, samples, and notes.
- ‡ Sample is already part of said Normal Data Set and is not selectable.

The second option in this mode is to add the sample to a new Normal Data Set. This will forward the user to the Create Normal Data Set page with the sample selected and require the user to enter in the remaining required fields.

The final option is the 'Include all job samples' checkbox. This will only be available if all the samples in the job can be added to Normal Data Sets. If this is selected the workflows are the same as above except that all of the job's samples are added. Note, a single Normal Data Set can have a maximum of 20 samples.

Existing Normal Data Sets

This is the initial page that is loaded when accessing Normal Data Sets from the configuration drop down link.



This page shows all Normal Data Sets a user has the permission to view. It shows how the Normal Data Sets are configured and provides a list of possible actions.

If the user who is viewing the page is not the owner of Normal Data Sets, the only action that they will have is cloning the Normal Data Set.

If they are the owner or have admin privileges they will also be able to:

- Change the owner.
- Change the visibility.
- Delete it, if no jobs have been run using it.
- Modify it, if it has not been committed yet.
- Upgrade it, if it is deprecated.

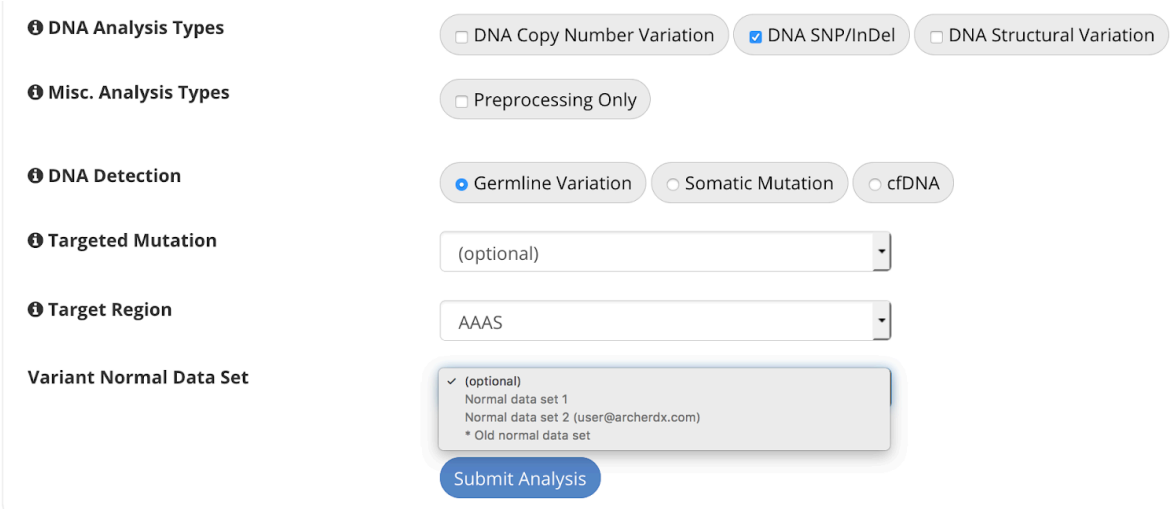
It is important to understand what changing the visibility of a Normal Data Set entails. If a user tries to change it to anything that is not Owner, they are giving others access to the sample ID and names that are contained within that Normal Data Set. The users who have access to the Normal Data Set then will be able to run jobs against it and clone the Normal Data Set if they want to make amendments. If giving others access to the Normal Data Set, a user must also remember to check that the panel that is associated with the Normal Data Set is also visible for those users. Otherwise they will not be able to use it when running jobs.

From this page the user can also get more information on the Normal Data Set by clicking on the Normal Data Set name. The details page will include:

- Samples included in the data set.
- Required settings and their values.
- Jobs that were run against the Normal Data Set.
- Watched folders that are using the Normal Data Set.

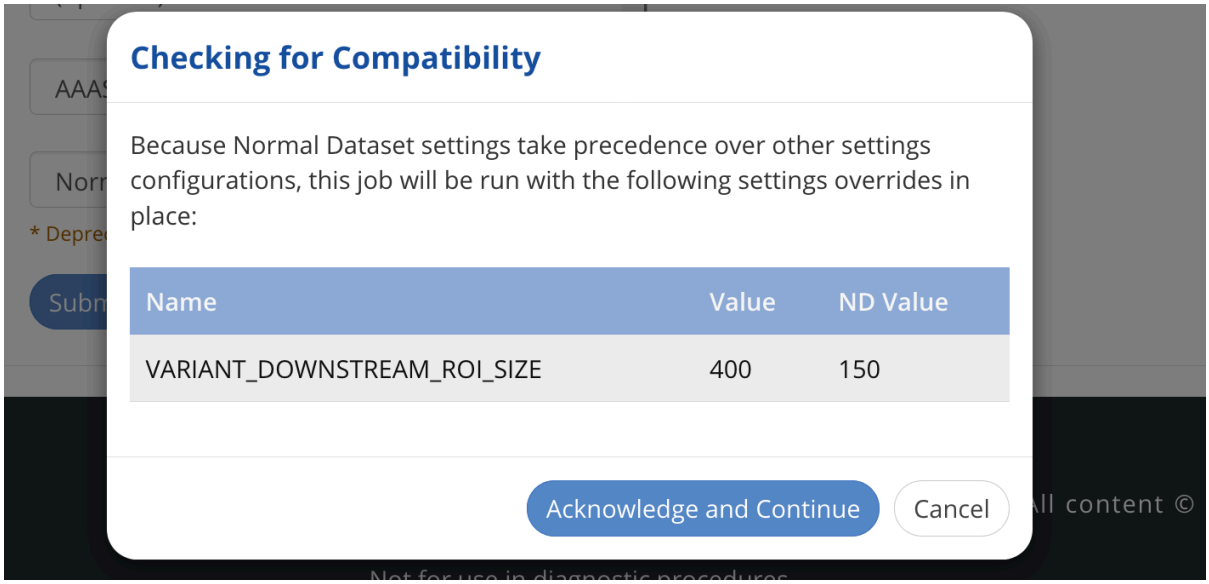
Performing Analysis with Normal Data Sets

Once the “Analysis Type” (SNP/InDel, DNA CNV or both) and the target region are selected, the Normal Data Set drop down will be populated. Only Normal Data Sets that the user has permission to see and match the analysis type, platform and panel, will be present.



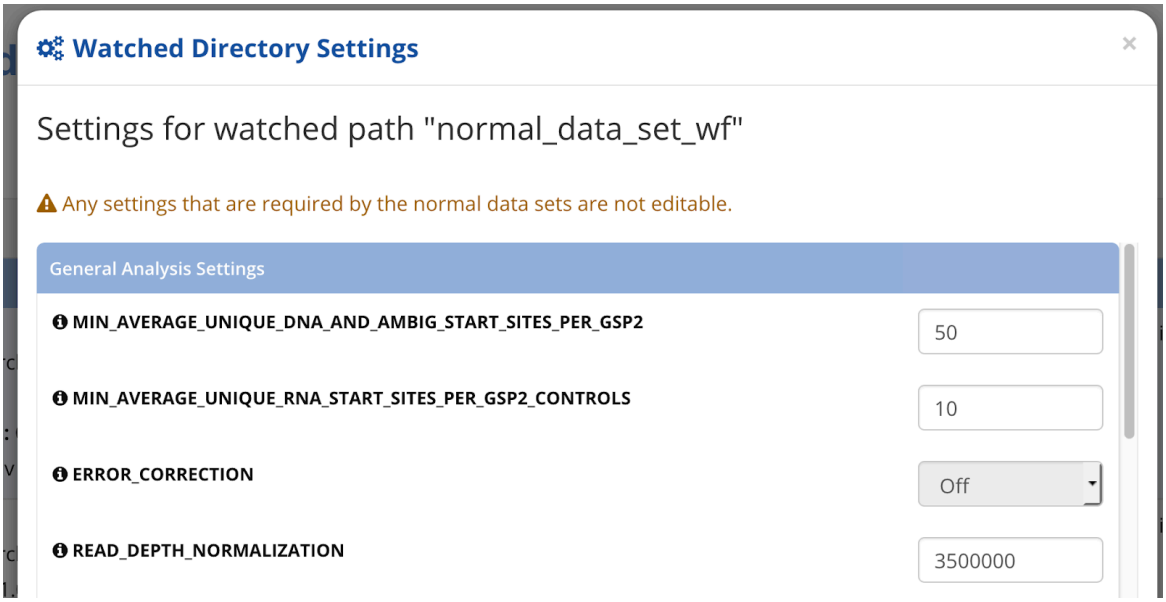
Normal data sets that are not deprecated are selectable. Ones with asterisks that precede their names are deprecated and must be upgraded prior to use.

When a user submits a job that uses a Normal Data Set, there is a chance that their settings do not match those that are required by the Normal Data Set. In those cases a warning will appear showing which user settings will be overridden by the Normal Data Set’s settings. The user must acknowledge that their settings will be overridden prior to job submission in these cases.



Watched Folders and Normal Data sets

Creating a watched folder with an associated Normal Data Sets works practically the same way as when performing an analysis as above except for how the settings are handled. When running an analysis, the system overrides the user settings so they match those of the Normal Data Set. In watched folders there is an additional constraint; watched folder settings that are required by the Normal Data Set are not editable after the watched folder is created.



Deprecated and Upgrading Normal Data Sets

If a Normal Data Set is deprecated that means that at least one of the samples that are part of it does not contain the required data to be used for new jobs. This might happen due to a software upgrade. If a Normal Data Set is deprecated it will be noted as such when looking at existing Normal Data Sets. An additional action of upgrade ( ) will also be available if the user is the owner.

There are a few additional things that happen on the system if a Normal Data Set is deprecated

- Any job that uses the Normal Data Set cannot be resubmitted nor cloned.
- No new jobs can use the Normal Data Set.
- Any existing watched folders that use the Normal Data Set are disabled.
- No new watched folders can be created that use the Normal Data Set.
- The Normal Data Set cannot be cloned.

When a user upgrades a Normal Data Set by clicking the upgrade action icon, all samples that are part of the Normal Data Set will be cloned and run on the system. These job names will be prefixed with “Normal Data Set upgrade rerun: ". Once all of the jobs finish successfully the Normal Data Set will then be upgraded and useable on the system again (see Job Setup and Management for general steps). Elements of this process that are unique to this pipeline are:

1. Select Analysis Type = “RNA Immune Repertoire”
2. Choose whether to calculate V-segment Hypermutation status (for top 5 most frequent clonotypes in IGH libraries)

Select the appropriate Target Region file (GTF file) for the specific Immunoverse panel used to generate your library (see *Experimental Features*)

Experimental features in Archer Analysis include advanced settings, additional pipelines, and supplementary functions that are not required for the standard use of the software, but may be of interest to expert users and for troubleshooting.

Accessing Experimental Features

Admin users by default have access to all experimental features. Admins can assign access to other users via the Edit User page (see the ‘Adding and/or Editing Users’ section above) and by either assigning the user to the ‘Experimental’ group, or by adding the ‘View experimental options’ permission:

Experimental Features for Job Set Up

These features are found on the Perform Analysis page:

DNA Target Coverage

Use this option in parallel with other DNA analysis types to report coverage metrics (i.e., depth at each base) for target regions of interest, as defined by a 'QC Target Regions' file (a BED file that can be uploaded via the 'Manage QC Target Regions' page).

Sample QC Only

Use this option under 'Misc. Analysis Types' to only generate data for the Read Statistics and Assay Targets pages (see Read Statistics and Assay Targets, respectively), i.e., samples will be processed only up to alignment and removal of off-target reads. This pipeline automatically enables the SAMPLE_QC setting.

Workflow Automation and Integrations

Watched Folders

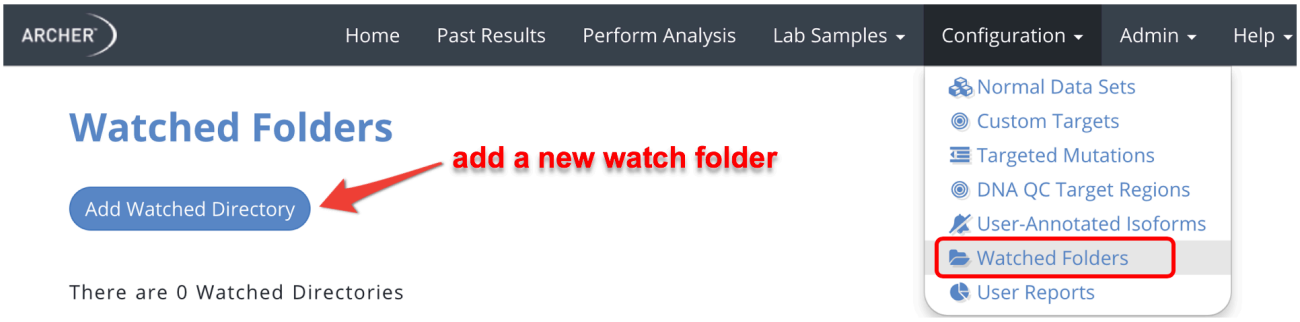
The Watched Folders mechanism allows the automatic execution of a predefined workflow whenever a set of FASTQ or BAM files are moved to a specified watched directory. There are two steps for the setup of the workflow automation:

1. Set up a watched folder in the web interface
2. Develop a script or procedure to move FASTQ or BAM files to the watch directory

Set Up a Watched Folder

To allow Archer Analysis to automatically execute a workflow when FASTQ or BAM files are placed in a watched folder, use the web interface to create a folder that is accessible to the sever:

Select the "**Configuration -> Watched Folders**" menu to reach the workflow automation page. This will also show you any watched folders that may already exist.



To add a new watch folder click the "**Add Watched Directory**" button. The dialog box is similar to the regular dialog box used to start a new analysis, with the exception of the "Folder Name" text box.

Type in the desired name of the watched directory in the Folder Name box (do not use spaces or symbols other than hyphen or underscore).

Configure assay type, platform and target region for the watch folder in the same way as it is done for a manual analysis. Each watched folder has to be unique and can only perform one specific workflow. Upon successful creation of the watched folder, you will now see it listed in the Watched Directories list.

Analysis Settings for the Executed Workflows

Within the existing Watched Folders, click on the Gear Icon in the Actions column to open the "Watched Folder Settings" Dialog:

Make any desired changes to these setting and click "Update Settings." Ensure that the analysis settings for the admin user are correct and as intended.

Removing a Watched Folder

To remove a watched directory, click the icon under Actions. A warning message will appear to ensure this was intended behavior. Removing a watched directory will NOT interfere or stop any currently running job, but will only avoid this directory from being considered for any future workflow execution.

Removing a watched directory will NOT delete the folder on the server but will only remove it from future consideration.

Removing Target Region Files That Are Used in Watched Folder Workflows

When a target region is defined as the target region for one or more automated workflow definitions, it cannot be removed until the watched directory itself is removed. This ensures that workflows that rely on the target region can be executed correctly.

The “Existing Custom Target” list will show if a target file is used in one or more workflow definitions:

To remove a target file, first remove all the analyses AND remove any watched directory definitions. After the removal, the delete icon will appear and the target region can be removed.

Develop Procedure/Script For the Movement of FASTQ or BAM Files

Once the watched folder has been created and the system has been setup to automatically execute a predefined workflow, a procedure or script needs to be developed to move, copy or link the required FASTQ or BAM files to the watched folder.

If the “archer” group has been created to allow only the archer_daemon and apache user to have write permission for the watched folder, ensure that whatever automation script is executed and actually fills the watch folder with the FASTQ or BAM files has the correct permissions for the watch folder.

How the System Knows When to Execute a Workflow

The workflow automation system continuously monitors each of the defined watched directories and looks for following files:

A file with the extension “[SAMPLE_NAME].completed”

A folder containing all the FASTQ files to be analyzed PLUS a file with the name “[FOLDER_NAME].completed” in the top level of the watched folder

In the first case, it will create a job in the Archer™ Analysis system and will run the workflow on the (pair of) FASTQ file(s) with the name [SAMPLE_NAME]*.fastq or [SAMPLE_NAME]*.fastq.gz (for the uncompressed and compressed version, respectively). In the second case it will create a job in the analysis system for ALL FASTQ files in the folder. The latter case will allow samples that are related (such as in CNV type of analysis, where there are case/control samples) to be run together.

Note the use of the ‘*’ (asterisk) wildcard symbol. In the case of Illumina paired-end sequence data, the names of the two FASTQ files are typically something like:

[SAMPLE_NAME]_R1_001.fastq

[SAMPLE_NAME]_R2_001.fastq

The suffixes “_R1_001” and “_R2_001” indicate the first and second read in the read pair data and the use of the ‘*’ character ensures that BOTH files are picked up for the workflow automation. For single read technologies such as Life Technologies™ Ion Torrent™ system or the Illumina single read libraries, create a file with the structure “[SAMPLE_NAME].completed”.

Example for a Job Containing a Single Sample

Here is an example of how to create a single sample for automatic execution. Content of the watch folder:

/var/www/analysis/watch_folder_ARR

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001.completed

Placing these three files in the watch directory will result in a job with the name “**BC-112_NA1473-FFPE_S11_L001**”

Always ensure that the FASTQ or BAM files are moved/copied to the watch folder BEFORE the “.completed” file is created. If the FASTQ or BAM files are NOT present, the job will produce an error.

Example for a Job Containing Multiple Samples

/var/www/analysis/watch_folder_ARR

|

+ - 4_Samples.completed

+ - 4_Samples

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

+ - BC-113_TriplePOS_S12_L001_R1_001.fastq

+ - BC-113_TriplePOS_S12_L001_R2_001.fastq

+ - BC-114_AmbionNEG_S13_L001_R1_001.fastq

+ - BC-114_AmbionNEG_S13_L001_R2_001.fastq

+ - BC-115_Water_S14_L001_R1_001.fastq

+ - BC-115_Water_S14_L001_R2_001.fastq

This will result in a job with the name "**4_Samples**" containing all 4 samples being analyzed at the same time. Note that the file "4_samples.completed" is located in the top level of the watch folder AND is created AFTER all the FASTQ files have been moved/copied/linked, to avoid errors in the job.

Fate of Files in the Watched Folder

Files (or links to files) that are placed in the watched directory will be removed from the watched directory and placed in the special directory "picked_up_files" (in the top level of the watched directory) where they will remain. The automation engine will create a symbolic link from this directory to the analysis directory (typically in "/var/www/analysis/[JOB_NUMBER]").

This is the structure of the watched directory BEFORE the files are picked up for analysis:

/var/www/analysis/watch_folder_ARR

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001.completed

This is the structure of the watch directory AFTER the files are picked up for analysis:

/var/www/analysis/watch_folder_ARR

|

+ - picked_up_files

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

The analysis folder/[JOB_NUMBER] contains symbolic links to the FASTQ files in the "picked_up_files" directory of the watch folder. It is imperative that the files in the "picked_ up_files" folder are NOT removed, since this will prevent the jobs from being re-run or cloned.

File Collisions

If the FASTQ files for a sample that had been previously analyzed by the workflow automation engine are placed in the watch directory again, the system will avoid the files from being overwritten in the "**picked_up_files**" directory by pre-pending the new files with the date and time and placing the new files in a special sub-directory of "**picked_up_files**" called "**collision_files**".

Here's an example of the situation BEFORE the sample BC-112 is run again:

/var/www/analysis/watch_folder_ARR

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001.completed

+ - picked_up_files

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

This is the situation after the files have been picked up again to be analyzed:

/var/www/analysis/watch_folder_ARR

|

+ - picked_up_files

|

+ - BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

+ - collision_files

|

+ - 01_06_2015_11_21_00_BC-112_NA1473-FFPE_S11_L001_R1_001.fastq

+ - 01_06_2015_11_21_00_BC-112_NA1473-FFPE_S11_L001_R2_001.fastq

The job will still be executed and the job will have the same name as the sample, but the FASTQ file name will be changed. The analysis/[JOB_NAME] folder will contain a link to the file in the **collision_files** directory.

S3 Watched Folders

Unlimited instances now have the option of using Amazon S3 as a watched folder input location.

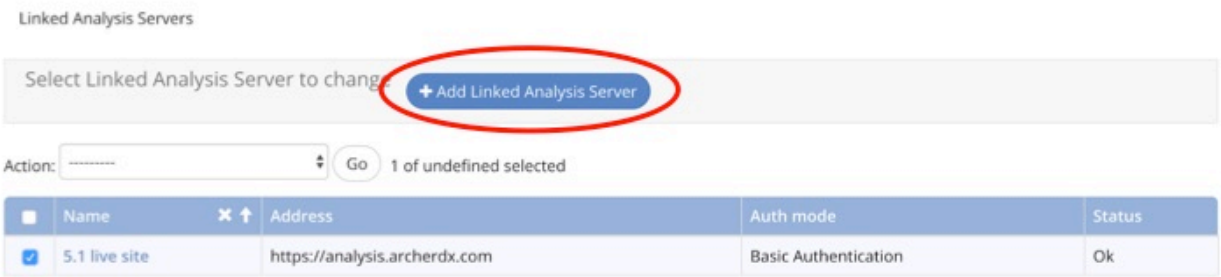
Additional setup of the **S3 Bucket** as well as notifying Archer of the **upload chunk size** that will be used to transfer files to the S3 bucket is required. Please contact tech@archerdx.com for additional details.

As a side note, customers can also use the REST API for job submission (SSL optionally supported) as an alternative to using watched folders. Please contact tech@archerdx.com for REST API documentation.

SSL-Enabled Linked Analysis Servers

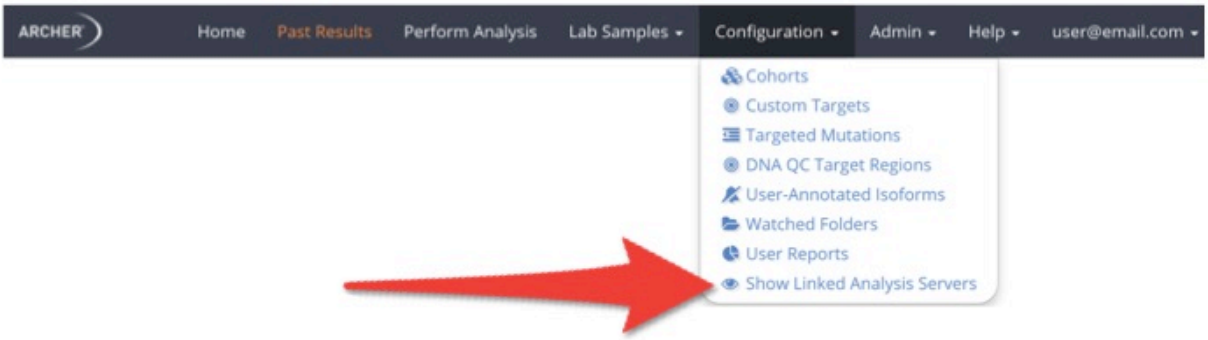
Linked Analysis Servers allow users to clone jobs between difference versions of instances of Archer Analysis. Admin users can access "Linked Analysis Servers" by selecting "**Admin**">"**Linked Analysis Servers**". Users can reference secure servers that have HTTPS enabled via the REST API using SSL.

Adding Linked Analysis Server

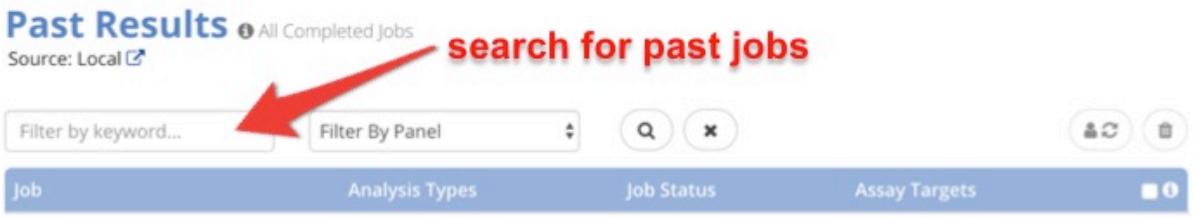


Enabling Linked Analysis Server

Admin users are given the permission to link Archer Analysis instances. From the Menu bar, select "**Past Results**" to access all the results that were run through the Archer Analysis pipeline. Once on the "**Past Results**" page, select "**Configuration**">"**Show Linked Analysis Servers**"



Under the "**Past Results**" header, Admin users will be able to see "**Source: Local ( )**" and change the server source to a different instance.



Once the source has been switched, users can search for jobs run on different instances of Archer Analysis.

Troubleshooting

This section contains information on dealing with problems that may arise during and immediately after the processing of jobs (for problems that arise during installation of the VM, please consult the **Archer Analysis VM Installation Guide**).

Job finished with job status "COMPLETED_ERROR"

This status indicates that one or more of the samples showed some sort of error during processing. The Summary Page may show one or more jobs with the message "Sample processed with errors".

If only one or a few of the jobs finished with errors, it is likely that the error happened during the processing of the sample. Click the "Processing log" link to see the log file to determine the location of the error. The easiest way to find the error is to search for the text "[ERROR]" in a text editor.

A common reason for failure of an individual job is running out of disk space:

```
[20:32:24] waterfall.sh [INFO] > Checking for required files...
[20:32:24] waterfall.sh [ERROR] > Can't find required file /var/www/analysis/1468/NA13-865_S10_L001_R2_001.molbar.trimmed.deduped.fastq
[20:32:24] waterfall.sh [ERROR] > Aborting.
/var/www/archer/analysis/shared_scripts/waterfall.sh: line 25: cannot create temp file for here-document: No space left on device
[20:32:24] run_waterfall_metrics_workflow [ERROR] > Non-zero exit code (1) from workflow, aborting.
```

This may occur if the /var/www/analysis directory on the virtual machine has reached its defined space. Deleting analyses that are no longer needed will free up some space. Alternatively, provision the virtual machine with more disk space. See the manual for the virtualization software for more information.

If the error is not pointing to an obvious solution, send the log files to tech@archerdx.com for support.

Job is stuck with job status "NEW"

This indicates that the queue manager software ("poller") is not running. The easiest way to restart the poller is to restart the complete virtual machine.

Alternatively, log into the Linux backend of the VM via the command line interface (see **Archer Analysis VM Installation Guide** for the credentials) and execute the following command in the **root home directory**:

$ ./restart-poller.sh

This should restart the poller and the job that was stuck in "NEW" should start running. To verify the poller is running, enter the following command:

$ ps –ef | grep poller

This should show a running python process called "poller.py"

```
[root@analysis ~]# ps -ef | grep poller
500       2164     1  0 Aug04 ?        00:00:46 python2.7 /var/www/html/archer_web/analysis/daemon/poller.py start
root     11199 11065  0 12:04 tty1     00:00:00 grep poller
[root@analysis ~]# _
```

If the poller is still not running, check the poller log file (stderr.poller) for more information.

If the VM was restarted it is likely that the job will show job status "COMPLETED_ERROR". Re-run the job by clicking the rerun icon ( ) to restart the job with the same settings.

Job is stuck with job status "HOLD"

This status shows that the job is still being processed although the processing step could have crashed. A job will show "HOLD" when it is processing the FASTQ files after upload.

If the job continues to show "HOLD" (for large files this could take up to an hour), restart the VM or restart the poller as shown in Job is stuck with job status "NEW"'.

Appendix

Glossary of General Terms Used in This Manual

GUI – Graphical User Interface. This is accessed via any modern web browser; Google Chrome is recommended.

VM – Virtual Machine. Archer Analysis includes all elements of a standalone computer, which can be run on top of virtualization software; VMware is recommended.

CLI – Command Line Interface. This is means for directly interacting with the underlying components of Analysis (Linux OS, etc.) by means of a terminal (console) window.

Intergenic – noncoding DNA regions between genes

Depth (DP) – The total high quality unique molecule depth covering the variant. High quality unique molecules are based on the Basequal and MAPQ filtering thresholds in user settings. Molecules not meeting the criteria will be filtered out of variant calling metrics.

Molecular Bins – Unique fragments/molecules from the original sample, as determined by use of molecular barcodes (MBCs) and deduplication. Molecular bins are defined as reads having the same random molecular barcode in the ligated adapter. This also requires the reads to have all the same start sites. The same MBC sequence could be seen in different start site locations.

MBC – Molecular Barcode in the ligated adapter.

GSP2 – The second Gene Specific Primer (GSP) used in the nested PCR of Anchored Multiplex PCR (AMP). It is the sequence that can be found in the final read output from the sequencer

Fusion Candidate (FC) – a fusion or isoform call that meets the minimum requirements specified in the User Settings.

Alternate Observations (AO) – this is the number of unique molecules supporting a variant call (alternative to the reference sequence, as defined by hg19)

Targeted Mutation File (TMF) – this file contains a list of specific SNVs and InDels, in Variant Call Format (VCF) v4.1, that is used with Targeted Mutation calling (Vision)

Format of the GTF Files

The Target Region file is a GTF file containing the names and locations of the primers. Gene Transfer Format (GTF) is described at this site:

http://genome.ucsc.edu/FAQ/FAQformat.html#format4

The ***gene_id*** field contains the name of the GSP2, and includes the name of the targeted gene.

The ***transcript_id*** field contains the name of the transcripts the target is derived from (can often be empty: “”).

FusionPlex panels have a ***function*** field containing the types of workflows that reads from this GSP2 should be processed with.

```
chr1    ArcherAssay GSP2    46290110    46290132    255 -   .   gene_id "MAST2_E_r_002_24_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    46295116    46295137    255 -   .   gene_id "MAST2_E_r_003_92_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    46425063    46425085    255 -   .   gene_id "MAST2_E_r_005_248_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    46463416    46463438    255 -   .   gene_id "MAST2_E_r_006_89_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    120465045   120465064   255 +   .   gene_id "NOTCH2_E_r_028_237_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    120465371   120465396   255 +   .   gene_id "NOTCH2_E_r_027_230_R_0_GSP2"; transcript_id ""; assay_type "Target";
chr1    ArcherAssay GSP2    120466580   120466602   255 +   .   gene_id "NOTCH2_E_r_026_229_R_0_GSP2"; transcript_id ""; assay_type "Target";
```

VariantPlex panels have a ***function*** field lists the applicable pipelines for each GSP2:

```
chr1    ArcherAssay GSP2    115251081   115251115   42  +   .   gene_id "NRAS_chr1_115251080_35_p_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115251135-
chr1    ArcherAssay GSP2    115251300   115251325   42  -   .   gene_id "NRAS_chr1_115251299_26_n_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115251135-
chr1    ArcherAssay GSP2    115252158   115252183   42  +   .   gene_id "NRAS_chr1_115252157_26_p_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115252169-
chr1    ArcherAssay GSP2    115252249   115252276   42  -   .   gene_id "NRAS_chr1_115252248_28_n_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115252169-
chr1    ArcherAssay GSP2    115256370   115256400   42  +   .   gene_id "NRAS_chr1_115256369_31_p_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115256400-
chr1    ArcherAssay GSP2    115256628   115256662   42  -   .   gene_id "NRAS_chr1_115256627_35_n_GSP2"; transcript_id "";function "SNV"; target_ROI "chr1:115256400-1152
chr1    ArcherAssay GSP2    115258636   115258658   42  +   .   gene_id "NRAS_chr1_115258635_23_p_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115258650-
chr1    ArcherAssay GSP2    115258791   115258819   42  -   .   gene_id "NRAS_chr1_115258790_29_n_GSP2"; transcript_id "";function "CNV,SNV"; target_ROI "chr1:115258650-
```

Format of the Targeted Mutation File (TMF)

The format of the targeted mutation file is a standard Variant Call Format (VCF) file, version 4.1, with some special INFO fields.

For more information about the format of VCF file, see the following web page:

https://samtools.github.io/hts-specs/VCFv4.1.pdf

The special INFO fields are:

- **Archer_Gene** – The gene the mutation is found in
- **Archer_MutationCDS** – The Coding DNA Sequence (CDS) change in HGVS format
- **Archer_MutationAA** – The Amino Acid change in HGVS format

An example of a TMF:

```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20170108
##reference=GRCh37
##INFO=<ID=Archer_CosmicID,Number=1,Type=String,Description="COSMIC Mutation Database identifier. Provided by Archer Analysis">
##INFO=<ID=Archer_Gene,Number=1,Type=String,Description="Gene name. Provided by Archer Analysis">
##INFO=<ID=Archer_MutationCDS,Number=1,Type=String,Description="Mutation in the CDS. Provided by Archer Analysis">
##INFO=<ID=Archer_MutationAA,Number=1,Type=String,Description="Amino Acid mutation in the protein. Provided by Archer Analysis">
##INFO=<ID=CNT,Number=1,Type=Integer,Description="How many samples have this mutation">
##INFO=<ID=STRAND,Number=1,Type=String,Description="Gene strand">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
chr1    11174395    COSM1185313 A   T   .   .   Archer_Gene=MTOR;STRAND=-;Archer_MutationCDS=c.7280T>A;Archer_MutationAA=p.L2427Q;CNT=4
chr1    11177096    COSM1662881 C   T   .   .   Archer_Gene=MTOR;STRAND=-;Archer_MutationCDS=c.6981G>A;Archer_MutationAA=p.M2327I;CNT=3
chr1    11177096    COSM4413485 C   A   .   .   Archer_Gene=MTOR;STRAND=-;Archer_MutationCDS=c.6981G>T;Archer_MutationAA=p.M2327I;CNT=1
```

When creating a custom TMF, it is recommended to use a provided VCF file as a guide to ensure the proper format of the file is maintained. Otherwise, Analysis will not be able to validate the file, and upload will fail. For example, it is important that the REF column contains the correct hg19 reference base. It is also important that the INFO header fields are included.

Read Stats & QC Metrics

Overall Sample Read Statistics

Total Fragments – Total number of read (pairs) that were present in the original FASTQ file.

Fragments with Complete Adapter – Total number of reads that contained the common region and 8mer molecular barcode. This also includes reads that may later be removed because they were too short.

Number of Reads After Trimming Adapter – Total number of reads (pairs) that are greater than 35bp after trimming the adapters, the common region and 8mer barcode.

Total Fragments (#) – Total number of fragments (read pairs) that pass the initial quality filter

Mapped (#/%) – Total number and percentage of fragments (read pairs) that map to the genome. Percentage is compared to the total number of fragments that pass the initial quality filter. [NUMBER OF MAPPED READS/NUMBER OF TOTAL READS x 100= % MAPPED]

Passed Alignment File – Percentage of fragments (read pairs) that pass the Alignment Score filter compared to the total number of fragments that map. Alignment Score setting used can be found in the Analysis Settings page. Default Setting = 30

On Target (%) – Percentage of fragments (read-pairs) with an alignment to the region of the genome targeted by the fragment's GSP2, compared to the total number of fragments that pass the mapping filter. Low % on target can be caused by the selecting wrong panel GTF, promiscuous primers or ribosomal RNA.

Molecular Bins – Molecular bins are defined as reads having the same random molecular barcode (8-mer) in the ligated adapter.

Average Molecular Bins per GSP2 – The total number of molecular bins divided by the total number of target GSP2.

Unique Start Sites – Unique start sites are defined as read 1 having a unique start site.

Average Unique Start Sites per GSP2 – The total number of reads with a unique start site divided by the total number of target GSP2 (Gene Specific Primer 2).

DNA Reads (#/%) – Total reads that likely come from a DNA source; reads that at least partially map to introns. Reads will be put into this category if they:

- Do not have a break in alignment to hg19 of more than 100 bp (aka a split)
- Include an intron region that must be at least 10% of the read length

RNA Reads (#/%) – Total reads that likely come from an RNA source (reads that span exon-exon splice junctions). Reads will be put into this category if they contain a split (a greater than 100 bp gap in alignment to hg19, which will occur when introns are spliced out)

Ambiguous Reads (#/%) – Total reads that map completely within exons: reads that are ambiguous as to their source. Reads will be put into this category if they do not have enough information to be placed in either of the other categories. (For example if a read does not contain a split, but also does not contain an intron region that is 10% of the length of the read.)

Mean Length (bp) – Apparent average fragment length as calculated by the mean of the total number of mapped reads. This will cap the apparent fragment length to 2 x [READ LENGTH] and should be considered an underestimate of the actual fragment length

Median Length (bp) – Median fragment length as calculated by the median of the total number of mapped reads. This is a better estimate of the actual fragment length since the fragment length is capped at 2X [READ LENGTH] and the median is less sensitive to this capping

Fusion/Isoform Metrics

Reads (#/%) – The number and percentage of unique reads supporting this gene fusion based on the molecular barcode. Only reads spanning the breakpoint are considered to support the fusion. Paired reads where both reads completely cover only one of the genes are NOT considered as supporting a gene fusion. The percentage is calculated in reference to the total number of reads covering this target, including wild-type transcripts. The read must extend 5 bp past GSP2 in order for the fusion to be called on the GSP2 side & this read must read 30 bp into the fusion partner for this to be called.

Start Sites – The number of unique reads supporting this gene fusion based on the unique start sites. Only reads spanning the breakpoint are considered to support the fusion. Paired reads where both reads completely cover only one of the genes are NOT considered as supporting a gene fusion.

Breakpoint – The hg19 chromosomal breakpoint locations for the fusion as deduced from the RNA. This does NOT represent the exact breakpoint at the DNA level. (Coordinates at the start of chr1 are placeholders for unaligned bases).

Output From 'User Defined File Download List'

There is a large amount of data obtained from the "**User Defined File Download List**" link. This includes output from all pipelines run in a job (including both raw text output, BAMs and plots in the form of PDFs and image files) as well as log files. The output is encapsulated in a single folder, which is named after the job ID. This folder is organized as follows:

Sample-Level Data

Sample-level files feature filenames with the name of the original FASTQ file (for READ1 if paired-end), but without the ".fastq" extension.

- Processing and error log files for each sample (numbered according to the order it was assigned at sample processing). These are found directly in the main folder
- BAM files – these are found directly in the main folder
- VCF files – these are found in the **VCFs** subfolder
- Custom flat text files in TSV, CSV, SVG and HTML formats summarizing read data and final results (from each pipeline). These are found directly in the main folder, as well as in the **counts** subfolder.
- Plots of various read metrics, in PDF or PNG image format. These are found in the **plots** subfolder.

Job-Level Data

- BED file representing the Archer panel used to process the job. This is found directly in the main folder.
- Custom flat text files in TSV, CSV, and VCF format summarizing data from individual text files above. These are found in the **summaries** subfolder.
- Text files from CNV analysis. These are found directly in the main folder
- Text files RNA expression analysis. These are found in the **summaries** subfolder.

Full Results File (full_results.txt)

The results of the Archer Analysis Software for each sample is provided in a single file containing all key read metrics, as well as results from the Fusion/Isoforms and DNA Anomalies pipelines.

***_full_results.txt**

Below is the result of the sample with the FASTQ file "1305_S12_L001_R1_001.fastq".

1305_S12_L001_R1_001_full_results.txt

The format of the file is a simple KEY VALUE pair, where the KEY and VALUE are separated by a TAB character:

SAMPLE_NAME = The name of the sample, which is the name of the FASTQ file, without the ".fastq" extension. For paired end reads it is the concatenated names of the original FASTQ files separated by an underscore (_).

FUSION_QC_FILTER = The results of the Fusion Quality Control filter. Will be PASS if it passes all QC filter settings or will indicate one or more values, indicating a potential issue with the library. Specific for the RNA Fusion type of analysis

VARIATIONS_QC_FILTER = The results of the Variations Quality Control filter. Will be PASS if it passes all QC filter settings or will indicate one or more values, indicating a potential issue with the library. Specific for the SNP/InDel types of analysis

MOLBAR_TOTAL_NUM_READS = Total number of read (pairs) in the FASTQ file used as input

MOLBAR_READS_WITH_CORRECT_COMMON_REGION = Total number of reads with the correct "common region" (no mismatches allowed)

MOLBAR_FRACTION_OF_TOTAL = Fraction of reads that have a perfect or near perfect common region:

(MOLBAR_READS_WITH_CORRECT_COMMON_REGION + MOLBAR_READS_WITH_CLOSE_COMMON_REGION)/ MOLBAR_TOTAL_NUM_READS

JUNK_PERCENT = The percentage of total reads that appear to be random sequence and not the result of the Archer AMP technology. Defined as the fraction of reads that don't align to the genome

MOLBAR_TOTAL_NUM_READS = FRAGMENTS_ALIGNED_FILTERED) / MOLBAR_TOTAL_NUM_READS

FRAGMENT_TOTAL = The total number of reads (pairs) for this sample

FRAGMENT_ALIGNED = The total number of reads (pairs) that align with the human genome (hg19)

FRAGMENT_ALIGNED_PERCENT = 100*FRAGMENT_ALIGNED / FRAGMENT_TOTAL

FRAGMENT_ALIGNED_FILTERED = Total number of reads (pairs) that align and also pass the mapping quality/alignment score filtering step

FRAGMENT_ALIGNED_FILTERED_ON_TARGET = Number of reads that have at least one of the pair aligned, pass alignment filtering, and is on-target with at least 1 base pair overlap

FRAGMENT_ALIGNED_FILTERED_ON_TARGET_PERCENT = 100*FRAGMENT_ALIGNED_FILTERED_ON_TARGET / FRAGMENT_ALIGNED_FILTERED

FRAGMENT_ALIGNED_FILTERED_OFF_TARGET = Number of reads where at least one of the pair is determined to be off-target and the other is not on-target

FRAGMENT_ALIGNED_FILTERED_OFF_TARGET_PERCENT = 100*FRAGMENT_ALIGNED_FILTERED_OFF_TARGET / FRAGMENT_ALIGNED_FILTERED

READ_n_TOTAL = The total number (n) of reads for this sample.

READ_n _ALIGNED = The total number of reads that align to the reference genome

READ_n _ALIGNED_PERCENT = The percentage of aligned reads, compared to the number of total reads (READ_n_TOTAL)

READ_n_ALIGNED_FILTERED = Number of reads with a mapping quality ≥ 35. Reads with a mapping quality below that value are removed from the analysis

READ_n_ALIGNED_FILTERED_ALONE_ON_TARGET = The number (n) of reads (after alignment score filtering) that align with at least 1 base on the define target region (controls and fusion gene candidates).

READ_n_ALIGNED_FILTERED_ALONE_ON_TARGET_PERCENT = The percentage of on-target reads, relative to the Filtered Total Molecules (READ_n_ALIGNED_FILTERED)

READ_n_ALIGNED_FILTERED_ALONE_OFF_TARGET = The number of read n reads (after alignment score filtering) that align and do not fall inside the defined target region (controls and fusion gene candidates)

READ_n_ALIGNED_FILTERED_ALONE_OFF_TARGET_PERCENT = The percentage of off-target reads is relative to the Filtered Total Molecules (READ_n_ALIGNED_FILTERED)

READ_n_ALIGNED_FILTERED_EITHER_ON_TARGET = The number of read1 or read2 reads (after alignment score filtering) that align with at least 1 base on the define target region (controls and fusion gene candidates).

READ_n_ALIGNED_FILTERED_EITHER_ON_TARGET_PERCENT = The percentage of on-target reads is relative to the Filtered Total Molecules (READ_n_ALIGNED_FILTERED)

READ_n_ALIGNED_FILTERED_EITHER_OFF_TARGET = The number of read1 or read2 reads (after alignment score filtering) that align and do not fall inside the defined target region (controls and fusion gene candidates)

READ_n_ALIGNED_FILTERED_EITHER_OFF_TARGET_PERCENT = The percentage of off-target reads is relative to the Filtered Total Molecules (READ_n_ALIGNED_FILTERED)

UNIQUE_* = The keys described above prefixed with UNIQUE have the same definition, but refer to the reads AFTER de-duplication (either alignment- or Molecular Barcoded-based)

FRAGMENT_EXON = The number of fragments (read-pairs) that contain a split alignment

FRAGMENT_INTRON = The number of fragments (read-pairs) that have at least 10% of the read covering an intron

FRAGMENT_EXON_PERCENT = 100*FRAGMENT_EXON/(FRAGMENT_EXON+FRAGMENT_intron). NOTE: a read could be counted TWICE if it both covers an exon AND an intron for at least 10%

FRAGMENT_MEAN_LENGTH = The average (deduced) length of the fragment (read pair)

FRAGMENT_MEDIAN_LENGTH = The median (deduced) length of the fragment (read pair)

RNA_FRAGMENT_MEAN_LENGTH = The average (deduced) length of the fragment (read pair) that is unambiguously categorized as RNA (span exon/exon boundaries)

RNA_FRAGMENT_MEDIAN_LENGTH = The median (deduced) length of the fragment (read pair) that is unambiguously categorized as RNA (span exon/exon boundaries)

DNA_FRAGMENT_MEAN_LENGTH = The average (deduced) length of the fragment (read pairs) that is unambiguously categorized as DNA (reads that cross from exon into intron without being split)

DNA_FRAGMENT_MEDIAN_LENGTH = The median (deduced) length of the fragment (read pair) that is unambiguously categorized as DNA (reads that cross from exon into intron without being split)

TOTAL_DNA_READS = Reads that are unambiguously categorized as DNA reads (reads that cross from exon into intron without being split)

TOTAL_RNA_READS = Reads that are unambiguously categorized as RNA reads, (reads that span an exon/exon boundary)

TOTAL_AMBIG_READS = Reads that fall completely inside an exon or intron and therefore cannot be categorized as either DNA or RNA

RNA_FRAGMENT_MEAN_LENGTH = The average length of fragments for reads classified as RNA reads (spanning at least two exons)

RNA_FRAGMENT_MEDIAN_LENGTH = The median length of fragments for reads classified as RNA reads (spanning at least two exons)

DNA_FRAGMENT_MEAN_LENGTH = The average length of fragments for reads classified as DNA reads (read that runs from exon to its next intron)

DNA_FRAGMENT_MEDIAN_LENGTH = The median length of fragments for reads classified as DNA reads (read that runs from exon to its next intron)

AMBIG_FRAGMENT_MEAN_LENGTH = The average length of fragments for reads classified as Ambiguous reads

AMBIG_FRAGMENT_MEDIAN_LENGTH = The median length of fragments for reads classified as Ambiguous reads

UNIQUE = The six keys described above that contain the string UNIQUE have the same definition, but refer to the reads AFTER de-duplication (Molecular Barcoded)

READS_PER_TARGET_n_TARGET_GENE = The name of the target region for which the data is relevant. n is the target number, starting at 0

READS_PER_TARGET_n_TYPE = The type of target. Can be either “TARGET” or “CONTROL”.

READS_PER_TARGET_n_READ_1 = Reads covering at least 1 bp of the target (GSP2) for read 1

READS_PER_TARGET_n_READ_2 = Reads covering at least 1 bp of the target (GSP2) for read 2 (for paired-end reads only)

READS_PER_TARGET_n_READ_EITHER = Reads covering at least 1 bp of the target for read 1 or read2

READS_PER_TARGET_n_READ_1_PERCENT = Percentage of the reads covering at least 1 bp of the target for read 1, compared to the total number of reads (READ_1_ALIGNED_FILTERED)

READS_PER_TARGET_n_READ_2_PERCENT = Percentage of the reads covering at least 1 bp of the target for read 2, compared to the total number of reads (READ_2_ALIGNED_FILTERED)

READS_PER_TARGET_n_READ_EITHER_PERCENT = Percentage of the reads covering at least 1 bp of the target for read 1 or read 2, compared to the total number of fragments (read pairs) (FRAGMENT_ALIGNED_FILTERED)

READS_PER_TARGET_n_RNA_READS = Number of reads classified as RNA reads (spanning two or more exons) that map to this GSP2

READS_PER_TARGET_n_RNA_READS_PERCENT = The percentage of the reads classified as RNA reads (spanning two or more exons) that map to this GSP2

READS_PER_TARGET_n_DNA_READS = Number of reads classified as DNA reads (reads that span a consecutive intron/exon boundary) that map to this GSP2

READS_PER_TARGET_n_DNA_READS_PERCENT = The percentage of the reads classified as DNA reads (reads that span a consecutive intron/exon boundary) that map to this GSP2

READS_PER_TARGET_n_NUCLEIC_ACID_READS = Number of total reads (DNA + RNA + Amb.) that map to this GSP2

READS_PER_TARGET_n_NUCLEIC_ACID_READS_PERCENT = The percentage of total reads (DNA + RNA + Amb.) that map to this GSP2

READS_PER_TARGET_n_AMBIG_READS = Number of reads classified as AMBIGUOUS reads (reads falling entirely inside a single exon) that map to this GSP2

READS_PER_TARGET_n_AMBIG_READS_PERCENT = The percentage of the reads classified as AMBIGUOUS reads (reads falling entirely inside a single exon) that map to this GSP2

DNA_FRAGMENT_GSP2_n_MEAN_LENGTH = The mean length of the fragments for reads classified as DNA reads (reads spanning a consecutive intron/exon boundary) that map to this GSP2

DNA_FRAGMENT_GSP2_n_MEDIAN_LENGTH = The median length of the fragments for reads classified as DNA reads (reads that span a consecutive intron/exon boundary) that map to this GSP2

RNA_FRAGMENT_GSP2_n_MEAN_LENGTH = The mean length of the fragments for reads classified as RNA reads (reads spanning two or more exons) that map to this GSP2

RNA_FRAGMENT_GSP2_n_MEDIAN_LENGTH = The median length of the fragments for reads classified as RNA reads (reads spanning two or more exons) that map to this GSP2

AMBIG_FRAGMENT_GSP2_n_MEAN_LENGTH = The mean length of the fragments for reads classified as AMBIGUOUS reads (reads falling entirely inside a single exon) that map to this GSP2

AMBIG_FRAGMENT_GSP2_n_MEDIAN_LENGTH = The median length of the fragments for reads classified as AMBIGUOUS reads (reads falling entirely inside a single exon) that map to this GSP2

UNIQUE_* = The keys described above prefixed with UNIQUE have the same definition, but refer to the reads AFTER de-duplication taking into account only reads with unique start sites

UNIQUE_START_SITES_* = The keys described above prefixed with UNIQUE_START_SITES have the same definition, but refer to the reads AFTER de-duplication taking into account only reads with unique molecular barcodes (bins)

TOTAL_UNIQUE_DNA_READS = Total number of unique DNA reads based on the molecular barcode (bins)

TOTAL_RAW_DNA_READS = Total number of DNA reads based before deduplication

TOTAL_UNIQUE_DNA_START_SITES = Total number of unique DNA reads based on the unique start sites

AVERAGE_UNIQUE_DNA_READS_PER_GSP2 = Average number of unique DNA reads per GSP2 based on the molecular barcode (bins)

AVERAGE_DNA_READS_PER_GSP2 = Average number of DNA reads per GSP2 based before deduplication

AVERAGE_UNIQUE_DNA_START_SITES_PER_GSP2 = Average number of unique DNA reads per GSP2 based on the unique start sites

TOTAL_UNIQUE_RNA_READS = Total number of unique RNA reads based on the molecular barcode (bins)

TOTAL_RAW_RNA_READS = Total number of RNA reads based before deduplication

TOTAL_UNIQUE_RNA_START_SITES = Total number of unique RNA reads based on the unique start sites

AVERAGE_UNIQUE_RNA_READS_PER_GSP2 = Average number of unique RNA reads per GSP2 based on the molecular barcode (bins)

AVERAGE_RNA_READS_PER_GSP2 = Average number of RNA reads per GSP2 based before deduplication

AVERAGE_UNIQUE_RNA_START_SITES_PER_GSP2 = Average number of unique RNA reads per GSP2 based on the unique start sites

TOTAL_UNIQUE_AMBIG_READS = Total number of unique ambiguous reads based on the molecular barcode (bins)

TOTAL_RAW_ AMBIG _READS = Total number of ambiguous reads based before deduplication

TOTAL_UNIQUE_ AMBIG _START_SITES = Total number of unique ambiguous reads based on the unique start sites

AVERAGE_UNIQUE_ AMBIG _READS_PER_GSP2 = Average number of unique ambiguous reads per GSP2 based on the molecular barcode (bins)

AVERAGE_ AMBIG _READS_PER_GSP2 = Average number of ambiguous reads per GSP2 based before deduplication

AVERAGE_UNIQUE_ NUCLEIC_ACID _START_SITES_PER_GSP2 = Average number of unique ambiguous reads per GSP2 based on the unique start sites

TOTAL_UNIQUE_ NUCLEIC_ACID _READS = Total number of unique total nucleic acid (DNA + RNA + Ambig) reads based on the molecular barcode (bins)

TOTAL_RAW_ NUCLEIC_ACID _READS = Total number of total nucleic acid (DNA = RNA + Amb.)reads based before deduplication

TOTAL_UNIQUE_ NUCLEIC_ACID _START_SITES = Total number of unique total nucleic acid (DNA + RNA + Amb.) reads based on the unique start sites

AVERAGE_UNIQUE_ NUCLEIC_ACID _READS_PER_GSP2 = Average number of unique total nucleic acid (DNA + RNA + Amb.) reads per GSP2 based on the molecular barcode (bins)

AVERAGE_ NUCLEIC_ACID _READS_PER_GSP2 = Average number of total nucleic acid (DNA +RNA + Amb.) reads per GSP2 based before deduplication

AVERAGE_UNIQUE_ NUCLEIC_ACID _START_SITES_PER_GSP2 = Average number of unique total nucleic acid (DNA + RNA + Amb.) reads per GSP2 based on the unique start sites

***_CONTROL** = The 24 keys described above with the PREFIX _CONTROL have the same definition but are limited to the control targets only.

FRAGMENT_GSP2_n_NAME = The name of the target region for which the data is relevant. n is the target number, starting at 0

FRAGMENT_GSP2_n_MEAN_LENGTH = The average (deduced) length of the fragment (read pair) for reads for this specific target [n]

FRAGMENT_GSP2_n_MEDIAN_LENGTH = The median (deduced) length of the fragment (read pair) for reads for this specific target [n]

READS_PER_TARGET_n_RNA_READS = The total number of RNA reads for this specific target [n]

AVERAGE_UNIQUE_RNA_READS_PER_GSP2 = The average number of reads classified as RNA reads (spanning two or more exons) that map to a GSP2.

COVERAGE_000_GENE = The coverage metrics for the bases in all targeted exons of the gene. “Summary” indicates the results are for all exons covered by GSP2’s, combined

COVERAGE_000_MIN_COV = The minimum coverage

COVERAGE_000_MAX_COV = The maximum coverage

COVERAGE_000_MEAN_COV = The mean coverage

COVERAGE_000_MEDIAN_COV = The median coverage

COVERAGE_000_PERCENT_10X_OR_GREATER = The percent of bases in the exons that have at least 10X coverage

COVERAGE_000_ PERCENT_100X_OR_GREATER = The percent of bases in the exons that have at least 100X coverage

COVERAGE_000_ PERCENT_1000X_OR_GREATER = The percent of bases in the exons that have at least 1000X coverage

COVERAGE_000_ PERCENT_BASES_GT_20_PRCT_MEAN = The percentage of bases that have coverage at least 20% of the mean coverage (COVERAGE_000_MEAN_COV)

COVERAGE_000_TOTAL_BASES = Total number of bases covered by the targeted exons

COVERAGE_000_ MIN_FOLD_CHANGE_70% = The ratio between the 70th percentile and the 30th percentile coverage value. This ratio is an indication of the “evenness of coverage”

COVERAGE_000_ MIN_FOLD_CHANGE_90% = The ratio between the 90th percentile and the 10th percentile coverage value. This ratio is an indication of the “evenness of coverage”

FC_n_GSP2 = Name of the target region associated with this Fusion Candidate (FC). n represents the fusion candidate number (starting at 1)

FC_n_GENES = The two (or more) genes participating in this gene fusion.

FC_n_KNOWN_FUSION = Indicates if the gene fusion candidate is a KNOWN fusion or not. The Archer Quiver database is used as the source. Values can be TRUE or FALSE

FC_n_INTRON_EXON_FUSION = Indicates if the gene fusion candidate is a fusion between an exon and an intron, often a sign of a false positive finding. Values can be TRUE or FALSE

FC_2_MISPRIMING_BASED_OFF_TARGET = Indicates if the two fusion partners share significant sequence similarity which is often caused by the primer mispriming of the not originally targeted gene. Values can be TRUE or FALSE

FC_n_ANNOTATION_1 = Annotation of the first fusion partner. Format: [GENE_NAME]([STRAND])|[exon|intron]:[0-9]*|[CHROM]:[STARTOFCONSENSUS],[CHROM]:[BREAKPOINT]

FC_n_ANNOTATION_2 = Annotation of the second fusion partner. Format: [GENE_NAME]([STRAND])|[exon|intron]:[0-9]*|[CHROM]:[STARTOFBREAKPOINT],[CHROM]:[STARTOFSECONDBREAKPOINT]

FC_n_BARCODE_ID = The identifier used in the consensus FASTA file. This identifier links the Fusion Candidate to the “molbar” identifier, used in the FASTA consensus sequence and BED file. i.e. If the FC_n_BARCODE_ID is 2, the identifier used in the consensus FASTA file is “2(GENE1:GENE2)_molbar_nn” where nn is some random number and (GENE1:GENE2) indicate the two gene fusion partners

FC_n_R1_COUNT = The number of filtered, non-redundant read 1 reads supporting this fusion candidate.

FC_n_R2_COUNT = The number of filtered, non-redundant read 2 reads supporting this fusion candidate.

FC_n_EITHER_R1_OR_R2 = The number of filtered, non-redundant read 1 OR 2 reads supporting this fusion candidate.

FC_n_BOTH_R1_AND_R2 = The number of filtered, non-redundant fragments (read-pairs) supporting this fusion candidate.

FC_1_UNIQUE_START_SITES = The number of reads with unique start sites that support the fusion candidate

FC_n_PROTEIN_TRANSLATION_x_GENE_BEFORE = The name of the fusion partner on the 5' side of the breakpoint (on RNA)

FC_n_PROTEIN_TRANSLATION_x_TRANSCRIPT_BEFORE = The (s) transcript for the 5' gene fusion partner (separated by a forward slash (/) if there is more than one) that could result in the sequence provided in the SEQUENCE field for this protein translation. If multiple transcripts are listed, it means that the AA sequence for these transcripts was the same.

FC_n_PROTEIN_TRANSLATION_x_GENE_AFTER = The name of the fusion partner at the 3' side of the break point (on RNA)

FC_n_PROTEIN_TRANSLATION_x_TRANSCRIPT_AFTER = The (s) transcript for the 3' gene fusion partner (separated by a forward slash (/) if there is more than one) that could result in the sequence provided in the SEQUENCE field for this protein translation. If multiple transcripts are listed, it means that the AA sequence for these transcripts was the same.

FC_n_PROTEIN_TRANSLATION_x_INFRAME = Indicates if the 3' fusion transcript is in frame with the 5' fusion partner (Yes or No)

FC_n_PROTEIN_TRANSLATION_x_SEQUENCE = The deduced AA sequence for this fusion partner/transcript combination. When out of frame and no stop codon is found before the last codon of the transcript, the last AA is the last AA of the last exon.

WT_n_GSP2 = Name of the target region associated with this Wild Type (non-fusion) isoform (WT)

WT_n_NOVEL = Indicates the isoform is a NOVEL isoform or not. Novel isoforms are defined as those isoforms that have non-consecutive exon/intron numbering in the annotation, suggesting exon skipping events

MOLBAR_TOTAL_NUM_READS = The total number of read (pairs) in the FASTQ files (only provided when using the Molecular Barcode based de-duplication)

MOLBAR_READS_WITH_CORRECT_COMMON_REGION = The number of read (pairs) in the FASTQ file that contain the common-region following the random Molecular Barcode. Only those reads with a perfect match to the common region sequence pass this filter.

BAM Files

All BAM files are after trimming and deduplication - by default Archer Analysis does not output raw alignments. For Illumina data, [R1|R2] indicates that there are separate files for each read direction; if omitted, the R1/R2 files are merged into an R1 file with a specific extension.

For all runs, the following BAMs will be generated:

***[R1|R2].molbar.trimmed.deduped.bam** - separate R1/R2 files for read alignments. Note for fusions, reads will be split (soft-clipped) to the respective areas of the genome they've been aligned to.

***molbar.trimmed.deduped.merged.bam** - this is the merged R1/R2 bam, sorted by coordinate. Use this to see all reads/alignments in the sample.

***molbar.trimmed.deduped.sorted.merged.bam** - this is the same as the R1/R2 merged bam, except that it is sorted by query name rather than coordinate.

For variant (SNP/InDel) workflows, the following additional BAM will be generated:

***preprocessed.bam** - this is the file used for the variant calling workflow. InDels have been realigned, low quality reads have been filtered out, and InDel qualities have been added (if using somatic processing). Use this to see exactly what data was used by the variant callers when searching for variants.

VCF Files in "VCFs" Subfolder

Where "*" is a placeholder for the sample name (as given by the original fastq file name). Please note that users should not use these files for downstream analysis:

***.vcf.summary.tsv** – This file is the tab delimited final results from the variant calling pipeline. This file matches the data displayed in the variant grid with no filtering.

***.[freebayes|lofreq|vision].ann.vcf** – This is the annotated, filtered output from each selected variant caller. (Vision calls are not filtered, as all targeted mutations will have a call/depth reported.)

***.vision.vcf** – This is the unannotated vcf covering all of the targeted mutations.

***.molbar.trimmed.deduped.freebayes.orig.vcf** – this is the raw output from Freebayes for this sample.

***.molbar.trimmed.deduped.freebayes.vcf** – this is the same as freebayes.orig.vcf, except with variants removed based on User Settings:

- MIN_DEPTH_FOR_VARIANT_CALL
- MAPQ_THRESHOLD_FOR_VARIANT_CALL
- MIN_BASEQUAL_FOR_VARIANT_CALL
- MIN_ALLELE_FRACTION_FOR_VARIANT_CALL
- MIN_PHRED_QUAL_SCORE_FOR_VARIANT_CALL
- VARIANT_DOWNSTREAM_ROI_SIZE (Note, the primer-specific ROI, if defined in the panel's GTF file, will override this system-wide setting).

***.molbar.trimmed.deduped.freebayes.ann.vcf** – this is the same as freebayes.vcf, except with all annotations from variant grid added (i.e. from COSMIC, Clinvar, and Ensembl VEP)

***.molbar.trimmed.deduped.indel.freebayes.vcf** – this is a subset of freebayes.vcf, listing only InDels.

The same set of rules apply to VCFs originating from the Lofreq caller, as described for Freebayes VCFs above:

- ***.molbar.trimmed.deduped.lofreq.orig.vcf**
- ***.molbar.trimmed.deduped.lofreq.vcf**
- ***.molbar.trimmed.deduped.lofreq.ann.vcf**
- ***.molbar.trimmed.deduped.indel.lofreq.vcf**

The same set of rules apply to VCFs originating from the targeted hotspot caller (Vision), as described for Freebayes VCFs above (note, there is no 'InDel' file for hotspot):

- ***.molbar.trimmed.deduped.hotspot.vcf**
- ***.molbar.trimmed.deduped.hotspot.ann.vcf**

Additional files not needed for further analysis:

- ***.molbar.trimmed.deduped.targeted_variants_outside_ROI.vcf** - lists all variants from all orig.vcf files listed above (from all callers), that fell outside the Region of Interest (ROI), as defined by User Settings (VARIANT_DOWNSTREAM_ROI_SIZE), or in the panels GTF file (if applicable).
- ***.molbar.trimmed.deduped.snp.freebayes.vcf** - file not to be used currently.
- ***.molbar.trimmed.deduped.snp.lofreq.vcf** - file not to be used currently.

VCF Files in "summaries" Subfolder

Summary-All-Variants.vcf – Compilation of all variants for all samples in job from these files. For duplicate variants, hotspot is chosen by default, otherwise the highest quality score from either Freebayes or Lofreq determines which caller's representation of the variant is presented. Additionally, annotations from COSMIC, etc. (i.e. the data added to **ann.vcf** files is included):

- ***.molbar.trimmed.deduped.freebayes.orig.vcf**
- ***.molbar.trimmed.deduped.lofreq.orig.vcf**
- ***.molbar.trimmed.deduped.hotspot.orig.vcf**

In addition to data from these files above, this file has additional columns for each sample in the job, listing the key read support metrics, if the variant was called for that sample (otherwise, this column will list blank values, and will not be listed in that sample's individual **.vcf** or **ann.vcf** files).

CNV Files

Job-level CNV files are located in the main folder:

- **cnv_sample_sheet.csv** – details sample groups, and which samples were Tumor or Normal.
- **CNV_SampleMetrics_1.csv** – summary metrics for all samples in the job.

Sample-level CNV files are also located in the main folder ("*" is a placeholder for the sample name):

- ***. molbar.trimmed.deduped.UNIQUE_READ_EITHER.counts_CNV_data_summary.csv** – is a listing of all CNV data (at each primer) for the sample.
- **molbar.trimmed.deduped.UNIQUE_READ_EITHER.counts_CNV_data_summary.pdf** – is the plot of CNV data that is presented in the GUI, in the CNV tab of the **Job Details** page.

CNV calling results

- ***.CNV.gene.csv** – the file contains the copy number and call for each gene in the panel.
- ***.CNV.primer.csv** – the file contains the copy number and call for each primer in the panel.

RNA Expression Files

These RNA expression files are located in the main folder:

- **rna_expression_visualization.pdf** – Heat map of relative expression of all primers across samples in the job (this file is also found in the **plots** subfolder).
- **rna_expression_visualization_sample_sheet.tsv** – details which samples are identified as Tumor in the job.

These RNA expression files are located in the **summaries** subfolder:

- **Summary.expression_summary_table.txt**
- **Summary.normalized_gene_expression.tsv**
- **Summary.per_sample_columns_expression_summary_table.txt**

Immune Repertoire Files

Immune repertoire output contains the same basic read metrics output as described for other pipelines. However, the files relevant to the core pipeline output (i.e., from MiXCR) are located in a dedicated subfolder for each sample (where "*" is a placeholder for the sample name):

***.molbar.trimmed.deduped.Immune_Repertoire_Results**

Within this subfolder, three key files summarize MiXCR results:

***.molbar.trimmed.deduped.Immune_repertoire.ImmuneRepertoireSummary.IR_aggregated_results.tsv**

***.molbar.trimmed.deduped.Immune_repertoire.ImmuneRepertoireSummary.txt**

***.molbar.trimmed.deduped.Immune_repertoire.mixcr_results.txt**

The remainder of the contents of this subfolder are TSV, PDF and SVG files containing data on specific clonotypes, segment usage, as well as many of the histograms featured in the GUI (see *Error! Reference source not found.* for further details).

Variant Summary Grid Columns

All of the columns in the four categories below can be viewed in the variant summary grid, and used to filter and sort results (see *Filtering Results* for further details).

General Variant Info

Actions - Basic Actions that can be run on the variant: (1) view local history, (2) view comments, (3) view reads in JBrowse, (4) create variant report, and (5) view in external databases (gnomAD, dbSNP, 1000 genomes).

Classification - Classification of variant or structural rearrangement.

Report – Whether or not this variant is considered 'reportable'. Variants or structural rearrangements annotated as such can more easily be reported on and are made more visible in the user interface.

Artifact – Variants deemed likely to be a chemistry or sequencing artifact.

Total Observed (TO) – Number of times this variant or structural rearrangement has been observed on this system.

Total Reported (Rept) – Number of times a variant has been marked as reportable on this system.

Path – ACMG Guideline-specified germline category: Pathogenic.

LPath – ACMG Guideline-specified germline category: Likely Pathogenic.

Bngn – ACMG Guideline-specified germline category: Benign.

LBngn – ACMG Guideline-specified germline category: Likely Benign.

Unc – ACMG Guideline-specified germline category: Uncertain Significance.

Tier I – AMP Guideline-specified somatic category: Tier I - 'Variants with strong clinical significance'

Tier II – AMP Guideline-specified somatic category: Tier II - 'Variants with potential clinical significance'

Tier III – AMP Guideline-specified somatic category: Tier III - 'Variants of unknown clinical significance'

Tier IV – AMP Guideline-specified somatic category: Tier IV - 'Variants deemed benign or likely benign'

Germline (Germ) – Variant is deemed likely to be germline.

Artifact (Artf) – Counts Variants deemed likely to be a chemistry or sequencing artifact.

Community Total Observed (C TO) – Community-sourced. Number of times this variant or structural rearrangement has been observed on this system.

Community Total Reported (C Rept) – Community-sourced. Number of times a variant has been marked as reportable on this system.

Community Path (C Path) – Community-sourced. ACMG Guideline-specified germline category: Pathogenic.

Community LPath (C LPath) – Community-sourced. ACMG Guideline-specified germline category: Likely Pathogenic.

Community Bngn (C Bngn) – Community-sourced. ACMG Guideline-specified germline category: Benign.

Community LBngn (C LBngn) – Community-sourced. ACMG Guideline-specified germline category: Likely Benign.

Community Unc (C Unc) – Community-sourced. ACMG Guideline-specified germline category: Uncertain Significance.

Community Tier I (C Tier I) – Community-sourced. AMP Guideline-specified somatic category: Tier I - 'Variants with strong clinical significance'

Community Tier II (C Tier II) – Community-sourced. AMP Guideline-specified somatic category: Tier II - 'Variants with potential clinical significance'

Community Tier III (C Tier III) – Community-sourced. AMP Guideline-specified somatic category: Tier III - 'Variants of unknown clinical significance'

Community Tier IV (C Tier IV) – Community-sourced. AMP Guideline-specified somatic category: Tier IV - 'Variants deemed benign or likely benign'

Community Germline (C Germ) – Community-sourced. Variant is deemed likely to be germline.

Community Artifact (C Artf) – Community-sourced. Variants deemed likely to be a chemistry or sequencing artifact.

gnomAD AF – The frequency of the allele called at this locus, from gnomAD global population.

Source – The variant caller that is the source of this variant call: LoFreq, FreeBayes or Vision.

Core Read Metrics

- **Depth (DP)** – The total high quality unique molecule depth covering the variant. High quality unique molecules are based on the Basequal and MAPQ filtering thresholds in user settings. Molecules not meeting the criteria will be filtered out of variant calling metrics.
- **Deep Depth (DDP)** – The total sequence coverage at this position that are from deep (i.e. error-correctable) molecular bins.
- **Deep Allele Frequency (DAF)** – The allele fraction of the reads from deep (i.e. error-correctable) molecular bins that support the alternative allele (DAO/DDP).
- **Unique start sites DP (UDP)** – The total unique start sites covering the variant.
- **Unique Allele Fraction (UAF)** – Allele fraction based on unique counts (UAO / UDP).
- **Alternate Observations (AO)** – Total number of reads that support the variant.
- **Unique start site AO (UAO)** – Total unique start sites supporting the variant.
- **Unique Reference Observations (URO)** – Total number of unique start sites represented by all the reference reads that intersect this variant.
- **Allele Fraction (AF)** – The fraction of the reads supporting the alternative allele (AO/DP).
- **Deep-bin AO (DAO)** – The number of alternate observations (AO) for this variant call that are from deep (i.e., error-correctable) molecular bins.
- **Deep-bin Reference Observations (DRO)** – The number of reference observations (RO) for this location that are from deep (i.e., error-correctable) molecular bins.
- **Deep-bin Rest (DRE)** – The number of non-reference and non-alternative observations for this location that are from deep (i.e., error-correctable) molecular bins.
- **DEEP_SHALLOW_THRESHOLD** – The depth at which a molecular bin will be considered a 'deep' bin. Deep bins will be considered higher confident reads as they have been assembled from multiple PCR duplicates. Default is 5 or greater.

- **Shallow-bin Rest (SRE)** – The number of non-reference and non-alternative observations for this location that are from shallow (i.e., non-error-correctable) molecular bins.

Homopolymer Count (HRUN) – The maximum homopolymer length the variant resides in according to the reference sequence and alt sequence.

COSMICID – The COSMIC ID associated with this mutation.

DBSNPID – The DBSNP ID associated with this mutation.

FATHMM – The FATHMM call as annotated by COSMIC for this mutation.

MutationalStatus – Mutational status as annotated by COSMIC.

FunctionalStatus – Functional status determined via COSMIC and FATHMM information. They can either be confirmed somatic/germline mutations in publications annotated in COSMIC, predicted somatic/germline via the FATHMM algorithm, or unknown.

Transcript ID (Trans) – The transcript ID that is associated with the current annotation of the variant (as reported by Ensembl VEP).

Reference Observations (RO) – Total number of reads that support the reference allele.

Sample Strand Bias Probability (SSB pVal) – p-value that quantifies the significance of discordance between reads originating from each strand of the source material for a given variant. It differs from Seq Dir Bias Prob in that it measures input material strand bias and NOT strand bias due to systematic sequencing error. LOW p-values are indicative of strand specific aberrations in the source material. These include C>U deamination events often found in FFPE samples.

Seq Dir Bias Probability (SDB pVal) – p-value that quantifies the significance of discordance between reads aligning to the positive strand versus those aligning to the negative strand for a given variant. It differs from Sample Strand Bias Probability in that it measures sequencing strand bias and NOT strand bias present in the input material. LOW p-values indicate strand specific aberrations introduced by the sequencing instrument.

Quality Rating – A text representation of the Quality Score. Low (0 to 1), Med (2-49) or High (50<).

MapQ – Indicates whether the variant was found in reads with alignments above the MapQ threshold (High) or below (Low). May also be No coverage.

Fishers Exact for Deep vs Shallow bins P-value (FEDSP) – P-value associated with the Fisher exact test performed between the error-corrected (deep) and the non-error-corrected (shallow) molecular bins. This measures the significance of the difference in allele fraction between the EC & non-EC reads. If p-value is LOW, it means the ratios are unusually different, which in turn may mean the variant call (likely from the non-EC bins) is not real, and instead caused by background noise (BN).

Fishers Exact for Deep vs Shallow bins odds Ratio (FEDSR) – The odds ratio that is related to FEDSP; the AF of deep bins (error-corrected) vs shallow bins (not error-correctable) Deviation from 1 indicates discordance between the two types of bins, and suggest the called variant (likely from non-EC bins) may not be real.

Canonical – Indicates "YES" if this transcript is considered the canonical transcripts. The canonical transcript is defined as either the longest CDS, if the gene has translated transcripts, or the longest cDNA.

Codons – The three-letter sequence for the codon the variant is found in. The variant base is shown as a capital letter.

Exon – The exon that contains the variant (exon/total_exons)

Genomic Location - Start position of the variant call (1- based, closed notation)

Primer Alt Reads + – The total number of alternate observations (AO) that came from the positive strand of fragments from the same primer (used to calculate Sample Strand Bias).

Primer Alt reads - – The total number of alternate observations (AO) that came from the negative strand of fragments from the same primer (used to calculate Sample Strand Bias).

Primer Ref Reads + – The total number of reference observations (RO) that came from the positive strand of fragments from the same primer (used to calculate Sample Strand Bias).

Primer Ref reads - – The total number of reference observations (RO) that came from the negative strand of fragments from the same primer (used to calculate Sample Strand Bias).

Quality Score – The PHRED based quality score of the variant call, as reported by the caller.

Ref/Alt Allele – The reference allele (hg19) and the alternative allele (i.e. the variant called).

Seq Alt Reads + – The total number of alternate observations (AO) that came primers targeting the positive strand (used to calculate Seq Dir Bias).

Seq Alt reads - – The total number of alternate observations (AO) that came primers targeting the negative strand (used to calculate Seq Dir Bias).

Seq Ref Reads + – The total number of reference observations (RO) that came primers targeting the positive strand (used to calculate Seq Dir Bias).

Seq Ref reads - – The total number of reference observations (RO) that came primers targeting the negative strand (used to calculate Seq Dir Bias).

Symbol – The Gene symbol for the gene located at this position (empty if variant is found in the intergenic region)

Type – Type of variant detected. Can be Single Nucleotide Polymorphism (SNP), Insertion or Deletion (InDel) or Complex, if more than one position is different

Variant Call – The variant call for this location. 0/0 represents a homozygous reference call. 0/1 and 1/1 represent heterozygous and homozygous alternative allele calls, respectively. For somatic mutations there can be 4 fields (i.e. 0/1/1/1) since for somatic mutations a ploidy of 4 is assumed)

Variant Name – (Targeted Mutation calling only) The variant name consists of the gene name and the amino acid mutation. This information is constructed from a set of special INFO fields in the provided mutations VCF file (Archer_Gene and Archer_MutationAA). See Format of the Targeted Mutation File (TMF) for further details.

Local Variant History

** See *SNP/InDel (RNA and DNA)* for further details on using these metrics**

Status - The report status (Not Reviewed, Reportable, etc.) that was selected for this variant, via the 'View History' option in the Actions column.

Reportable – The total number of times this variant was labeled as 'Reportable' (via the 'View History' tool in the Actions column) across all samples in the system.

Reviewed – The total number of times this variant was labeled as 'Reviewed' (via the 'View History' tool in the Actions column) across all samples in the system.

Total – The total number of times this variant was vetted (via the 'View History' tool in the Actions column) across all samples in the system.

Other – The total number of times this variant was vetted (via the 'View History' tool in the Actions column) across all samples in the system.

Advanced Statistical Metrics

Sample Strand Bias Ratio – The odds ratio of positive vs. negative strand (for Fisher's Exact test). Deviation from 1 indicates potential sample strand bias.

Has Sample Strand Bias – The determines if there is statistically significant asymmetry in the positive vs. negative strands, as determined by opposing primers targeting those strands – indicative of deamination or PCR error. This value is YES if the p-value (i.e. Sample Strand Bias Prob) ≤ 0.05.

Seq Dir Bias Ratio – The odds ratio of positive vs negative strand (for Fisher's Exact test). Deviation from 1 indicates potential sequencing strand bias.

Has Seq Dir Bias – The determines if there is statistically significant asymmetry in the sequencing reads (R1 vs. R2) of single library fragments – indicative of sequencing error. This value is YES if the p-value (i.e. Seq Dir Bias Prob) ≤ 0.05.

Biological Significance

HGVSc – (From Ensembl VEP) The mutation at the DNA Coding Sequence (CDS) level, in the format from the Human Genome Variant Society (concatenated with the currently chosen transcript ID – See *SNP/InDel (RNA and DNA)* for further details on customizing the transcript ID).

HGVSp – (From Ensembl VEP) The mutation at the amino acid (i.e., protein) level in the format, from the Human Genome Variant Society (concatenated with the currently chosen transcript ID – See *SNP/InDel (RNA and DNA)* for further details on customizing the transcript ID).

Clinical Significance – The clinical relevance of the variant according to the CLINVAR database.

Variant disease name – The disease name this variant could have clinical relevance for according to the CLINVAR database.

Consequence – The calculated coding consequence of the variation. See the Ensembl VEP page for more information about the various classes of consequence.

SIFT – SIFT predicts whether an amino acid substitution affects protein function. The amino acid substitution is predicted to be damaging if the score is ≤0.05 and tolerated if the score is > 0.05 (note that is the opposite to the PolyPhen score, where higher scores are considered deleterious)

PolyPhen – PolypPhen predicts whether an amino acid substitution affects protein function. The PolyPhen score represents the probability that a substitution is damaging. Values nearer 1 are more confidently predicted to be deleterious (note that this is the opposite to the SIFT score, where lower scores are considered deleterious).

MA Match – YES/NO value indicating if the detected variant matches any of the sub-population minor alleles

Intra-Job Allele Fraction Outlier P-value (Intra Job AF Outlier P Value) – The probability this mutation was due to background noise (BN), as estimated across all samples in the same job. A low p-value indicates the AF of the called variant is significantly outside the BN, providing confidence the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Intra-Job Background Noise (Intra Job BN) – The estimated background noise of this position in the genome, as estimated across all samples in the same job. This is estimated by performing allele fraction outlier detection on all samples in the job at this position. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Minimal Detectable Allele Fraction (Intra Job MDAF) – This is the lowest actual allele fraction (AF) that the variant could be at in the original sample, assuming binomial probability distribution, as estimated across all samples in the same job. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical AF would be at least this value 95% of the time.

Intra-Job Minimal Detectable Allele Fraction 95% (Intra Job 95MDAF) – The allele fraction (AF) at which a variant can be called with high statistical power, as estimated across all samples in the same job. In contrast to MDAF, if the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Alternate Observations Conservative (Intra Job AOC) – The AO associated with the lowest actual allele fraction that the variant could be at in the original sample, assuming a binomial probability distribution (see AFC), as estimated across all samples in the same job. If the sample was prepped & sequenced many times, the empirical AO would be at least this value 95% of the time.

Intra-Job Allele Fraction Conservative (Intra Job AFC) – This is the lowest actual allele fraction (AF) that the variant could be at in the original sample, assuming binomial probability distribution, as estimated across all samples in the same job. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical AF would be at least this value 95% of the time.

Intra-Job Deep Background Noise (Intra Job DBN) – The estimated background noise of this position in the genome, as estimated across all samples in the same job using only unique molecules with a deep amplicon depth. This is estimated by performing allele fraction outlier detection on all

samples in the job at this position. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Deep Allele Fraction Outlier P-value (Intra Job DAF Outlier P Value) - The probability this mutation was due to deep background noise (DBN), as estimated across all samples in the same job using unique molecules with a deep amplicon depth. A low p-value indicates the DAF of the called variant lies significantly outside the DBN, and provides confidence the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Intra-Job Deep Minimal Detectable Allele Fraction (Intra Job DMDAF) – The allele fraction at which we would consider a variant significant (i.e., above the background noise, BN), as estimated across all samples in the same job. If the true AF in the sample is at least this, and this identical experiment were run multiple times, 50% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Deep Minimal Detectable Allele Fraction 95% (Intra Job D95MDAF) – The allele fraction (AF) at which a variant can be called with high statistical power, as estimated across all samples in the same job using only unique molecules with a deep amplicon depth. In contrast to MDAF, if the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Deep Alternate Observations Conservative (Intra Job DAOC) – The AO associated with the lowest actual allele fraction that the variant could be at in the original sample, assuming a binomial probability distribution (see AFC), as estimated across all samples in the same job using only unique molecules with a deep amplicon depth. If the sample was prepped & sequenced many times, the empirical AO would be at least this value 95% of the time.

Intra-Job Deep Allele Fraction Conservative (Intra Job DAFC) – This is the lowest actual allele fraction (AF) that the variant could be at in the original sample, assuming binomial probability distribution, as estimated across all samples in the same job using only unique molecules with a deep amplicon depth. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical AF would be at least this value 95% of the time.

Intra-Job Shallow Allele Fraction Outlier P-value (Intra Job SAF Outlier P Value) – The probability this mutation was due to shallow background noise (SBN), as estimated across all samples in the same job using only unique molecules with a shallow amplicon depth. A low p-value indicates the SAF of the called variant lies significantly outside the SBN, and provides confidence the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Intra-Job Shallow Background Noise (Intra Job SBN) – The estimated background noise of this position in the genome, as estimated across all samples in the same job using only unique molecules

with a shallow amplicon depth. This is estimated by performing allele fraction outlier detection on all samples in the job at this position. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Shallow Minimal Detectable Allele Fraction (Intra Job SMDAF) – The allele fraction (AF) at which a variant can be called with high statistical power, as estimated across all samples in the same job using only unique molecules with a shallow amplicon depth. In contrast to MDAF, if the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Shallow Minimal Detectable Allele Fraction 95% (Intra Job S95MDAF) – The allele fraction (AF) at which a variant can be called with high statistical power, as estimated across all samples in the same job using only unique molecules with a shallow amplicon depth. In contrast to MDAF, if the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called.

Intra-Job Shallow Alternate Observations Conservative (Intra Job SAOC) – The AO associated with the lowest actual allele fraction that the variant could be at in the original sample, assuming a binomial probability distribution (see AFC), as estimated across all samples in the same job using only unique molecules with a deep amplicon depth. If the sample was prepped & sequenced many times, the empirical AO would be at least this value 95% of the time.

Intra-Job Shallow Allele Fraction Conservative (Intra Job SAFC) – This is the lowest actual allele fraction (AF) that the variant could be at in the original sample, assuming binomial probability distribution, as estimated across all samples in the same job using only unique molecules with a shallow amplicon depth. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical AF would be at least this value 95% of the time.

Normal Data Sets Allele Fraction Outlier P-value (ND AF Outlier P Value) – The probability this mutation was due to background noise given the provided Normal Data Set (ND_BN) and taking all consensus reads into account. A low p-value indicates the AF of the called variant lies significantly outside the ND_BN, and provides confidence that the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Normal Data Sets Background Noise (ND BN) – The estimated background noise of this position in the genome given the provided Normal Data Set and taking all consensus reads into account. This is estimated by performing allele fraction outlier detection on all samples in the job at this position.

Normal Data Sets Minimal Detectable Allele Fraction (ND MDAF) – The allele fraction at which we would consider a variant significant (i.e., above the background noise, ND_BN) given the provided Normal Data Set and taking all consensus reads into account. If the true AF in the sample is at least this, and this identical experiment were run multiple times, 50% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Minimal Detectable Allele Fraction 95% (ND 95MDAF) – The allele fraction (AF) at which a variant can be called with high statistical power given the provided Normal Data Set and

taking all consensus reads into account. In contrast to ND_MDAF, if the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Alternate Observations Conservative (ND AOC) – The AO associated with the lowest actual allele fraction that the variant could be at in the original sample, assuming a binomial probability distribution (see AFC), as estimated across all samples in the same job. If the sample was prepped & sequenced many times, the empirical AO would be at least this value 95% of the time.

Normal Data Sets Allele Fraction Conservative (ND AFC) – This is the lowest actual allele fraction (AF) that the variant could be at in the original sample, assuming binomial probability distribution, as estimated across all samples in the same job. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical AF would be at least this value 95% of the time.

Normal Data Sets Deep Allele Fraction Outlier P-value (ND DAF Outlier P Value) – The probability that this mutation was due to background noise given the provided Normal Data Set (ND_DBN) and unique molecules with a deep amplicon depth into account. A low p-value indicates the DAF of the called variant lies significantly outside the ND_DBN, and provides confidence that the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Normal Data Sets Deep Background Noise (ND DBN) – The estimated background noise of this position in the genome given the provided Normal Data Set and taking only 'deep' (composed of three or more raw reads) consensus reads into account. This is estimated by performing allele fraction outlier detection on all samples in the job at this position.

Normal Data Sets Deep Minimal Detectable Allele Fraction (ND DMDAF) – The deep allele fraction at which we would consider a variant significant (i.e., above the background noise, ND_DBN) given the provided Normal Data Set and taking only 'deep' (composed of three or more raw reads) consensus reads into account. If the true DAF in the sample is at least this, and this identical experiment were run multiple times, 50% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Deep Minimal Detectable Allele Fraction 95% (ND D95MDAF) – The allele fraction (DAF) at which a variant can be called with high statistical power given the provided Normal Data Set and taking only 'deep' (composed of three or more raw reads) consensus reads into account. In contrast to ND_DMDAF, if the true DAF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Deep Alternate Observations Conservative (ND DAOC) – The DAO associated with the lowest actual deep allele fraction that the variant could be at in the original sample given the background noise of the Normal Data Set and taking only 'deep' (composed of three or more raw reads) consensus reads into account. This assumes a binomial probability distribution (see ND_DAFC). If the sample was prepped & sequenced many times, the empirical DAO would be at least this value 95% of the time.

Normal Data Sets Deep Allele Fraction Conservative (ND DAFC) – This is the lowest actual deep allele fraction (DAF) that the variant could be at in the original sample, assuming a binomial probability distribution using the background noise of the Normal Data Set and taking into account only 'deep' (composed of three or more raw reads) consensus reads. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical DAF would be at least this value 95% of the time.

Normal Data Sets Shallow Allele Fraction Outlier P-value (ND SDAF Outlier P Value) – The probability this mutation was due to background noise given the provided Normal Data Set (ND_SBN) and taking only unique molecules with a shallow amplicon depth into account. A low p-value indicates that the SAF of the called variant lies significantly outside the ND_SBN, and provides confidence that the mutation is real. For each individual variant, outlier detection metrics will be reported if another sample has coverage over the variant. In addition, if the variant is not included in the targeted mutation file, it must also have 3 or more Deep Alternate Observations (DAO) for outlier detection metrics to be reported. Variants in the targeted mutation file do not have a DAO requirement for outlier detection metrics to be reported.

Normal Data Sets Shallow Background Noise (ND SBN) – The estimated background noise of this position in the genome given the provided Normal Data Set and taking only 'shallow' (composed of less than three raw reads) consensus reads into account. This is estimated by performing allele fraction outlier detection on all samples in the job at this position.

Normal Data Sets Shallow Minimal Detectable Allele Fraction (ND SMDAF) – The shallow allele fraction at which we would consider a variant significant (i.e., above the background noise, ND_SBN) given the provided Normal Data Set and taking only 'shallow' (composed of less than three raw reads) consensus reads into account. If the true SAF in the sample is at least this, and this identical experiment were run multiple times, 50% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Shallow Minimal Detectable Allele Fraction 95% (ND S95MDAF) – The allele fraction (SAF) at which a variant can be called with high statistical power given the provided Normal Data Set and taking only 'shallow' (composed of less than three raw reads) consensus reads into account. In contrast to ND_SMDAF, if the true SAF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant.

Normal Data Sets Shallow Alternate Observations Conservative (ND SAOC) – The SAO associated with the lowest actual shallow allele fraction that the variant could be at in the original sample given the background noise of the Normal Data Set and taking only 'shallow' (composed of less than three raw reads) consensus reads into account. This assumes a binomial probability distribution (see ND_SAFC). If the sample was prepped & sequenced many times, the empirical SAO would be at least this value 95% of the time.

Normal Data Sets Shallow Allele Fraction Conservative (ND SAFC) – This is the lowest actual shallow allele fraction (SAF) that the variant could be at in the original sample, assuming a binomial probability distribution using the background noise of the Normal Data Set and taking into account only 'shallow' (composed of less than three raw reads) consensus reads. It correlates to the lower bounds of the 95% confidence interval, which is in turn based on the depth of reads at the genomic location of concern. If the sample were prepped & sequenced many times, the empirical SAF would be at least this value 95% of the time.

Minimum Outlier P-value (Min Outlier P Value) – The minimum value of Intra-job AF, SAF, DAF and Normal Data Set AF, SAF, DAF p-values if a Normal Data Set was selected. If any test shows a statistically significant outlier, then it can be considered a significant variant. See the individual test definitions for details.

Allele Fraction Outlier P-value (AF Outlier P Value) – Representative AF Outlier P value for the sample. This will be the AF Outlier P-Value from the Normal Data Set if one was provided, otherwise this will be the intra-sample AF Outlier P-Value. If any test shows a statistically significant outlier, then it can be considered a significant variant. See the individual test definitions for details.

Minimal Detectable Allele Fraction 95% (95MDAF) – The allele fraction (AF) at which a variant can be called with high statistical power. If the true AF in the sample is at least this, and this identical experiment were run multiple times, 95% of the time there would be sufficient signal to capture this variant. Note, to show a value here, at least one other sample in the job must have the same variant called. This value will be the 95MDAF given the background of the normal data set if one was provided, otherwise it will be the 95MDAF of the intra-job comparison if outlier detection is on.

Summary of Icons

– low % GSP2 support for fusion/isoform candidate

– low confidence annotation on fusion/isoform candidate

– potential mispriming on fusion/isoform candidate

– passed all strong evidence criteria for fusion/isoform candidate

– fusion gene pair found in Quiver (but not exact breakpoint)

– exact breakpoint found in Quiver for fusion candidate

– possible transcriptional read-through event for fusion/isoform candidate

– known paralog in Ensembl (for fusion/isoform candidate)

– fusion with intronic content

– potential cross-contamination (for fusion candidate)

– expression imbalance support (for fusion candidate)

– negative support for fusion/isoform candidate (causing weak evidence binning)

– search or visualize read data (in JBrowse)

– turn visibility of variant summary grid column on/off

– click for informational pop-up

– menu with further options

– download files

– view analysis job settings

– clone a job

– failed QC/stop a running or queued job

– passed QC

– view/download log files

– view/print job or sample reports

– sample or mutation vet history dialogue window

– view Ensembl VEP data

– rerun a job

– Upgrade deprecated sample

Copyright and Trademark Notices

ArcherDX, Inc.

2477 55th Street, Suite 202
Boulder, CO 80301
303-357-9001
www.archerdx.com

CS001 (DOC-969)

Approved By:

Keli Gardner - Author

April 12, 2019 5:30 PM GMT 6b63a22f-52b8-4c0a-a3b9-fb52d9f07db4

Chloe Stabler - Document Control

April 17, 2019 8:31 PM GMT 9082ec98-3d87-4f44-ac46-b090edcfd141

Version History:

Author	Effective Date	Ver.	Status
Keli Gardner	April 17, 2019 12:00 AM GMT	0	Published