HELMHOLTZ MUNICH

Core Facility Bioinformatics and Statistics

Helmholtz Zentrum München

KaryoExplorer - Digital Karyotyping Pipeline

Illumina Demo Dataset Guide

Data Source and License

Contents

1. Required Software and Data

This guide provides step-by-step instructions for downloading and organizing the demo dataset for the KaryoExplorer - Digital Karyotyping Pipeline. The demo dataset consists of 36 samples and includes raw data files, manifest, cluster file, and sample sheet. The workflow described here is independent of official Illumina data preprocessing pipelines and reflects the data preprocessing steps performed for the KaryoExplorer pipeline.

Before beginning the data extraction process, ensure you have the following software and data components installed and downloaded. This chapter provides an overview of all required tools for the GenomeStudio software (Illumina) data extraction workflow.

1.1 Software Requirements

Operating System

GenomeStudio is a Windows-only application. Ensure your system meets the following requirements:

- **Operating System:** Windows 10 or Windows 11 (64-bit)
- **RAM:** 16 GB or more recommended for this dataset.
- **Hard Disk Space:** At least 50 GB free space

Windows Required

Important: GenomeStudio only runs on Windows operating systems. macOS and Linux users will need to use a Windows virtual machine, dual-boot setup, or dedicated Windows computer to complete this tutorial.

GenomeStudio Software v2.0.5

GenomeStudio is Illumina's primary software for analyzing microarray data, including genotyping, gene expression, and methylation analysis [1].

- **Version:** 2.0.5 (includes Genotyping v2.0.5 module)
- **Download URL:** `https://emea.support.illumina.com/downloads/genomestudio-2-0.html`
- **Release Date:** March 4, 2020
- **File Size:** 79 MB (ZIP archive)
- **Platform:** Windows 10/11 (64-bit) only
- **Components:**

  - GenomeStudio Software v2.0.5 Installer
  - Genotyping v2.0.5 module
  - Polyploid Genotyping v2.0.5 module
- **Documentation:** Release notes (PDF) available on the download page

Installation Note

GenomeStudio requires a valid license from Illumina. Contact Illumina support (`techsupport@illumina.com`) if you need assistance with licensing or installation.

PLINK Input Report Plug-in v2.1.4

The PLINK Input Report Plug-in is an essential add-on for GenomeStudio that enables export of genotyping data in PLINK format [2], [3], [4].

- **Version:** 2.1.4
- **Download URL:** `https://emea.support.illumina.com/downloads/genomestudio-2-0-plug-ins.html`
- **Purpose:** Export genotype data in PLINK-compatible formats (.ped, .map, .bed, .bim, .fam)
- **Installation:** Must be installed **after** GenomeStudio installation

Important

The PLINK plug-in must be installed before attempting to export data. Without this plug-in, you will not be able to generate the required output files for the Digital Karyotyping Pipeline.

2. Downloading the Demo Dataset

This chapter provides step-by-step instructions for downloading all required files for the Illumina Global Screening Array v4.0 demo dataset [5], [6]. The dataset consists of 36 samples and includes raw data files, manifest, cluster file, and sample sheet.

2.1 Overview of Required Files

The complete demo dataset consists of four main components:

1. **Demo Data (iScan Raw Data)**
   - 36 samples in iScan format
   - Raw intensity data files (.idat) [7], [8]
   - Red and Green channel files for each sample
   - Total: 72 .idat files (36 samples × 2 channels)
2. **Manifest File (BPM and CSV Format)**
   - Array design and probe information
   - Genomic coordinates (GRCh38/Build 38) [9]
   - SNP annotations and probe sequences
   - File: `GSA-48v4-0_20085471_D2.bpm`
   - File: `GSA-48v4-0_20085471_D2.csv`
3. **Cluster File (EGT Format)**
   - Genotype cluster positions
   - Reference population data for genotype calling
   - Version: v4 (48v4)
   - File: `GSA-48v4-0_20085471_D2_ClusterFile.egt`
4. **Sample Sheet (CSV Format)**
   - Sample metadata and identifiers
   - Links samples to .idat files
   - Specifies required product file versions
   - File: `GSA-48v4-0_D2_SampleSheet_Demo_36.csv`

2.2 Download Instructions

Before starting the data extraction process, you need to download the following files from Illumina's support website. All files should be downloaded and organized **before launching GenomeStudio software (Illumina)**.

Step 1: Download Demo Data

Navigate to the Global Screening Array Support Files page:

- **URL:** `https://emea.support.illumina.com/downloads/global-screening-array-support-files.html`

Download the following file (highlighted with red border in Figure 2.1):

1. **Infinium Global Screening Array v4.0 Demo Data (iScan)**
   - File format: ZIP archive
   - Size: 825 MB
   - Date posted: May 28, 2024
   - Contents: 36 samples with .idat files (Red and Green channels)
   - Total files: 72 .idat files (36 samples × 2 channels: Red and Green)

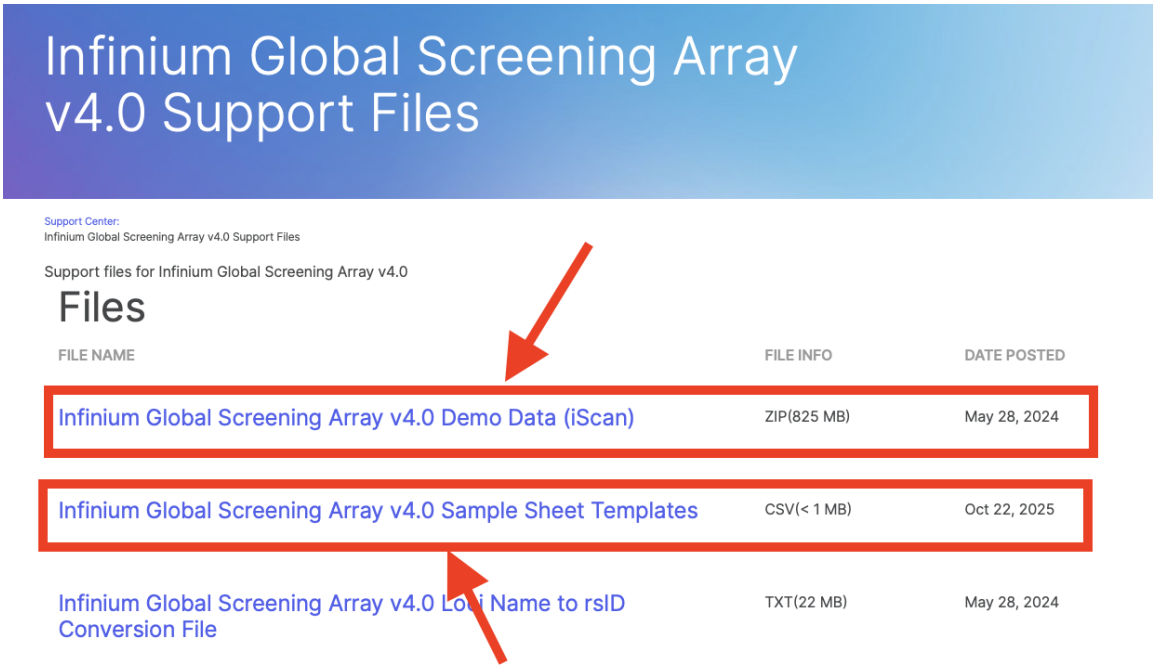

Figure 2.1: Illumina Global Screening Array v4.0 Support Files page. Download the files highlighted with red borders: Demo Data (iScan) and Sample Sheet Templates.

Step 2: Download Product Files and Sample Sheet

Navigate to the Global Screening Array Product Files page:

- **URL:** `https://emea.support.illumina.com/downloads/global-screening-array-v4-product-files.html`

Download the following files for **Build 38 (GRCh38)** (highlighted with red borders in Figure 2.2):

1. **Manifest File: GSA-48v4-0_20085471_D2.bpm**
   - File format: BPM
   - Size: 109 MB
   - Date posted: May 28, 2024
   - Genome Build: GRCh38 (Build 38)

- **Note:** Also available in CSV format (270 MB) for reference

2. **Cluster File: GSA-48v4-0_20085471_D2_ClusterFile.egt**
   - File format: EGT
   - Size: 48 MB
   - Version: 48v4-0
   - Date posted: May 28, 2024
3. **Sample Sheet: GSA-48v4-0_D2_SampleSheet_Demo_36.csv**
   - File format: CSV
   - Direct download: `https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium-gsa-with-gcra/GSA-48v4-0_20085471_D2.csv`
   - Contains: Sample identifiers, sentrix barcodes, and product file version information
   - Samples: 36 demo samples

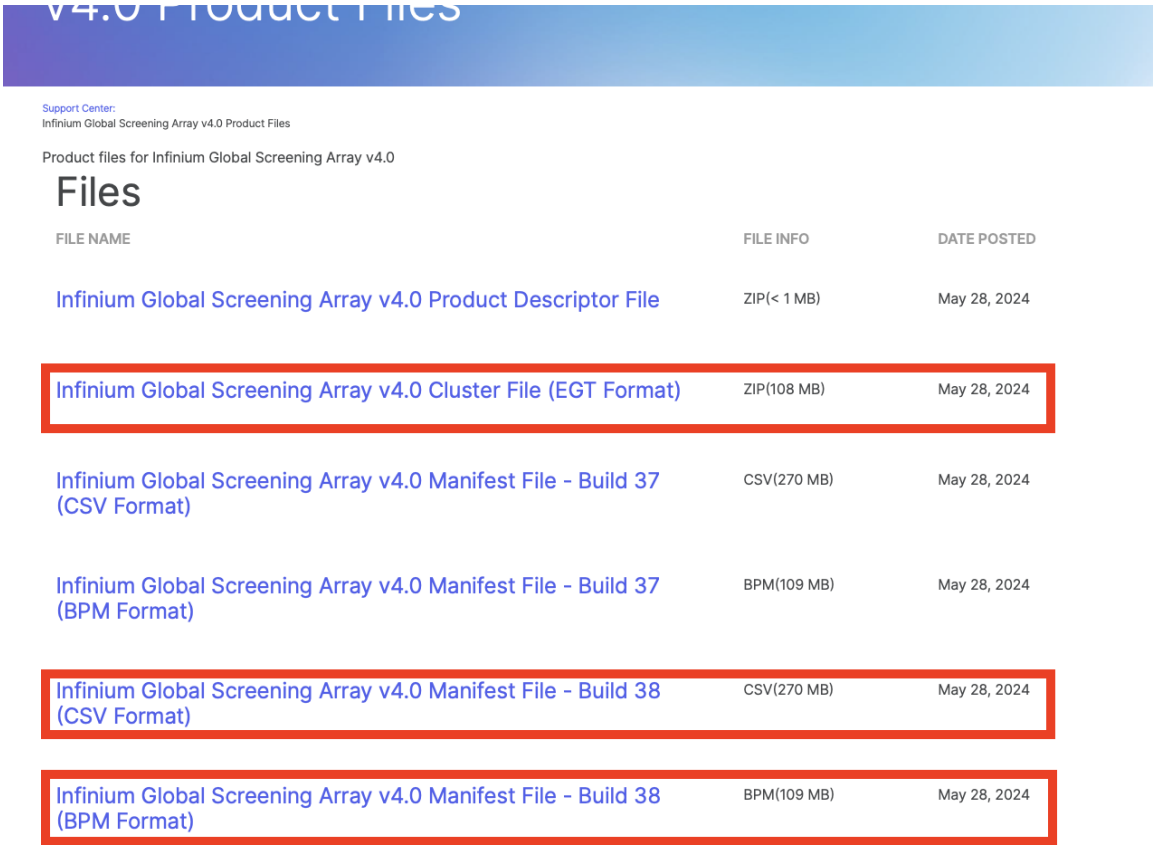

Figure 2.2: Illumina Global Screening Array v4.0 Product Files page. Download the files highlighted with red borders: Manifest File Build 38 (BPM Format), Cluster File (EGT Format), and Sample Sheet (CSV). The sample sheet can also be downloaded directly from the provided URL.

Important Note

Since the demo data is based on **GRCh38 (Build 38)**, ensure you download the corresponding manifest and cluster files for Build 38. Using mismatched genome builds will result in incorrect genomic coordinates and analysis errors. The sample sheet (`GSA-48v4-0_D2_SampleSheet_Demo_361.csv`) specifies the exact product file versions required for this dataset.

Step 3: Organize Downloaded Files

Create a dedicated directory on your Desktop to organize all files. Follow this structure:

1. Create the main directory:

```
Desktop/
`-- Infinium_Global_Screening_Array_v4.0/

```

2. Extract and organize the downloaded files under the **Infinium_Global_Screening_Array_v4.0** folder (the name of the folder is arbitrary, you can choose any name you want);

```
Infinium_Global_Screening_Array_v4.0/
|-- idats_Demo_36/
|   |-- 204939850001_R01C01_Red.idat
|   |-- 204939850001_R01C01_Grn.idat
|   |-- 204939850001_R01C02_Red.idat
|   |-- 204939850001_R01C02_Grn.idat
|   `-- ... (36 samples = 72 .idat files total)
|-- GSA-48v4-0_20085471_D2.bpm
|-- GSA-48v4-0_20085471_D2_ClusterFile.egt
`-- GSA-48v4-0_D2_SampleSheet_Demo_36.csv

```

2.3 Pre-Analysis Checklist

Before proceeding to create a genotyping project in Genome Studio, verify you have completed all the following steps:

Item	Status
Software Installation	
GenomeStudio v2.0.5 installed	☐
GenomeStudio license activated	☐
PLINK Input Report Plug-in v2.1.4 installed	☐
Data Downloaded	
Demo data downloaded (36 samples, 72 .idat files)	☐
Manifest file: GSA-48v4-0_20085471_D2.bpm	☐
Cluster file: GSA-48v4-0_20085471_D2_ClusterFile.egt	☐
Sample sheet: GSA-48v4-0_D2_SampleSheet_Demo_36.csv	☐
Data Organization	
All files organized in dedicated directory structure	☐
iScan raw data extracted from ZIP archive	☐
File paths are accessible and correct	☐

Table 2.1: Pre-analysis checklist - Complete before starting Genome Studio

Ready to Begin Genome Studio Analysis

Once all items in the checklist above are completed, you are ready to proceed with creating a genotyping project in Genome Studio. The next section will guide you through the step-by-step process of project creation, genotype calling, and data extraction.

3. Data Extraction with Genome Studio

This chapter provides detailed step-by-step instructions for extracting the required data files from GenomeStudio software (Illumina). These files are essential inputs for the Digital Karyotyping Pipeline and include genotype data, sample information, SNP annotations, and PLINK format files.

3.1 Overview of Required Output Files

The Digital Karyotyping Pipeline requires the following files to be extracted from Genome Studio:

1. **Full Data Table** (`Full_Data_Table.txt`)
   - Contains comprehensive genotyping data for all samples and SNPs
   - Includes LRR (Log R Ratio) and BAF (B Allele Frequency) values [10]
   - Critical for copy number variation (CNV) analysis [11]
2. **Samples Table** (`Samples_Table.txt`)
   - Contains sample-level quality control metrics
   - Includes call rates, heterozygosity [12], and other QC parameters
   - Used for sample quality assessment
3. **SNP Table** (`SNP_Table.txt`)
   - Contains SNP-level information and statistics
   - Includes genomic coordinates and allele frequencies
   - Used for marker quality control
4. **PLINK Files** (`.ped`, `.map`, `.bed`, `.bim`, `.fam`)
   - Standard genetic data format for downstream analysis [3], [4]
   - Generated using PLINK Input Report Plug-in [2]
   - Required for genotype-based analyses
5. **Manifest File** (`GSA-48v4-0_20085471_D2.csv`)
   - Array design information in CSV format
   - Already downloaded in section 2.2
   - Contains probe annotations and genomic coordinates

Important: Folder Structure

Make sure your folder structure is organized as follows before starting:

- **Sample Sheet:**
  `Infinium_Global_Screening_Array_v4.0/`
  `GSA-48v4-0_D2_SampleSheet_Demo_36.csv`
- **Data Repository:**
  `Infinium_Global_Screening_Array_v4.0/idats_Demo_36/`
  (contains .idat files)
- **Manifest Repository:**
  `Infinium_Global_Screening_Array_v4.0/`
  (contains .bpm and .egt files)

All three paths should point to locations within or at the `Infinium_Global_Screening_Array_v4.0/` directory.

3.2 Step 1: Creating a Genotyping Project

3.2.1 Launch Genome Studio

1. Open Genome Studio v2.0.5 from the Windows Start menu
2. Wait for the application to fully load
3. You will see the Genome Studio welcome screen

3.2.2 Start New Genotyping Project

1. Click on **File** → **New Project** (see Figure 3.1)
2. Select **Genotyping** from the project type options
3. The Genotyping Wizard will open (see Figure 3.2)

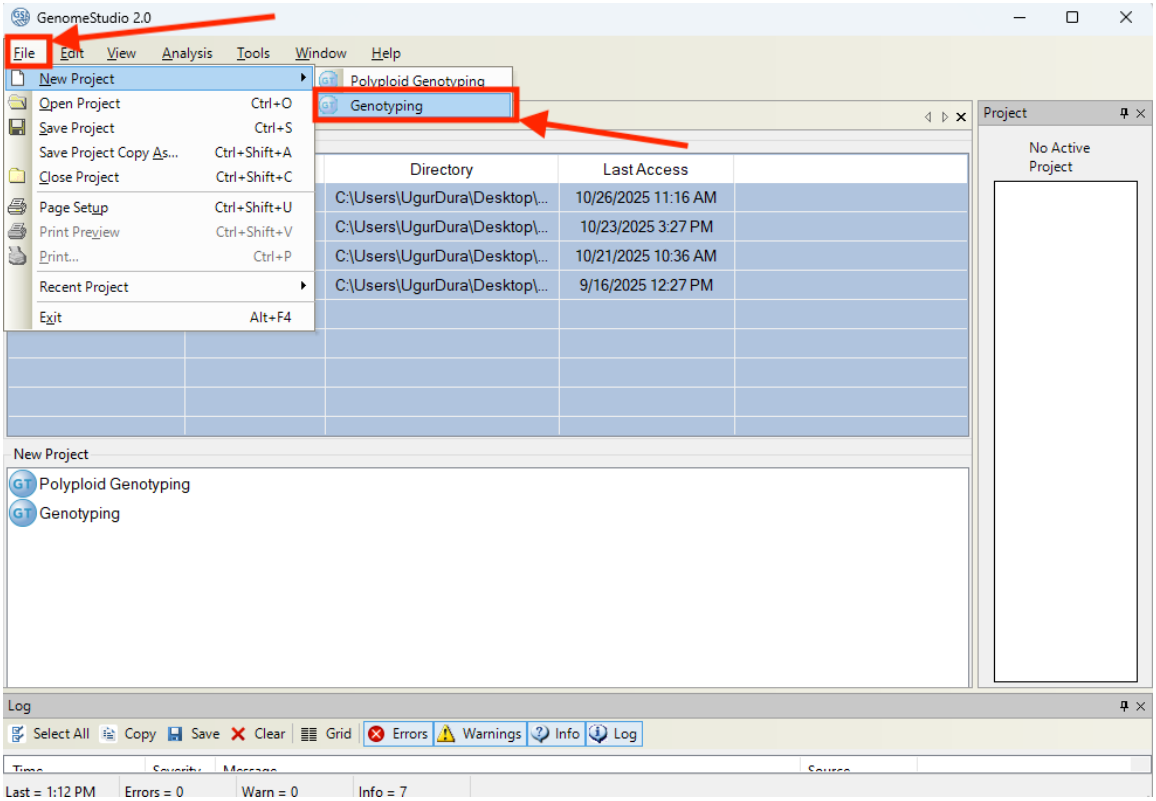

Figure 3.1: Creating a new genotyping project. Select File → New Project → Genotyping to launch the Genotyping Wizard.

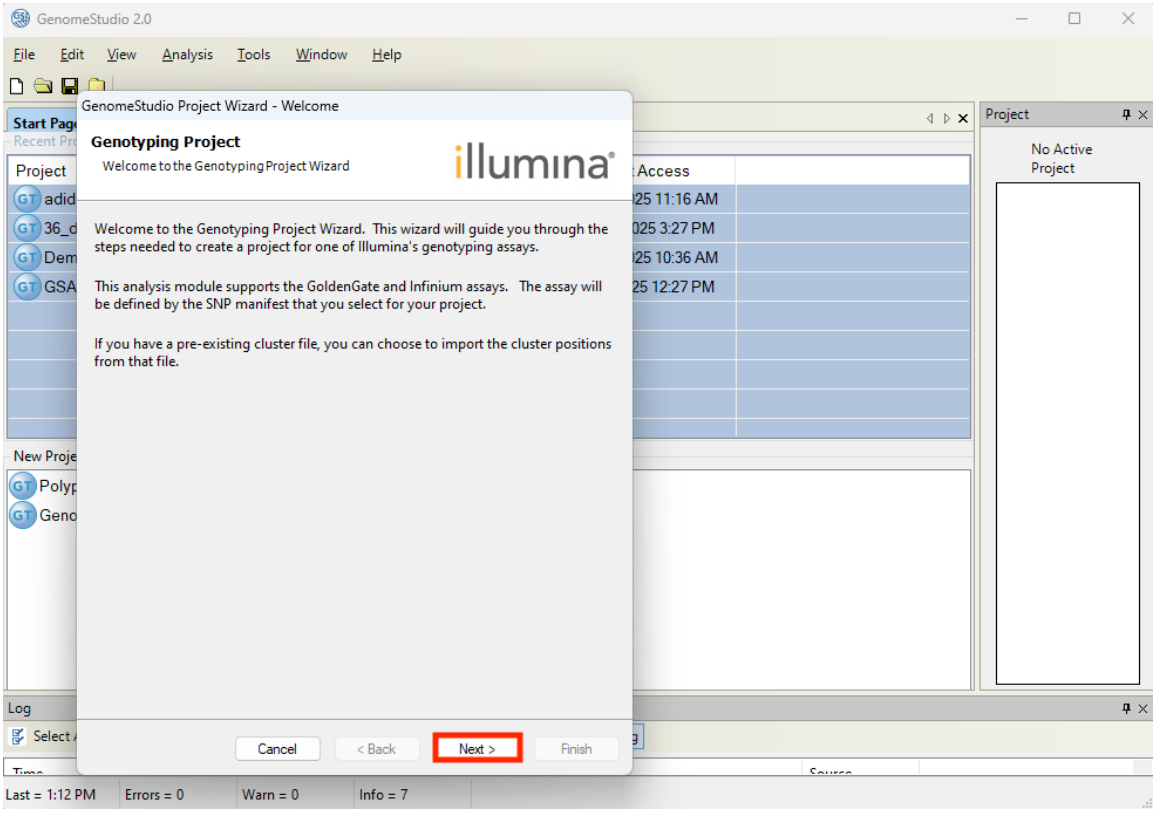

Figure 3.2: The Genotyping Wizard guides you through the project setup process, including project naming, sample sheet import, data repository selection, and manifest loading.

3.2.3 Name Your Project

The first step in the Genotyping Wizard is to name your project and choose where to save it.

1. Enter a descriptive project name (e.g., `Infinium_Global_Screening_Array`) (see Figure 3.3)
2. Choose a location to save the project file (.bsc)
3. Recommended location: `Infinium_Global_Screening_Array_v4.0/GenomeStudio_Project/`
4. Click **Next** to proceed to the next step

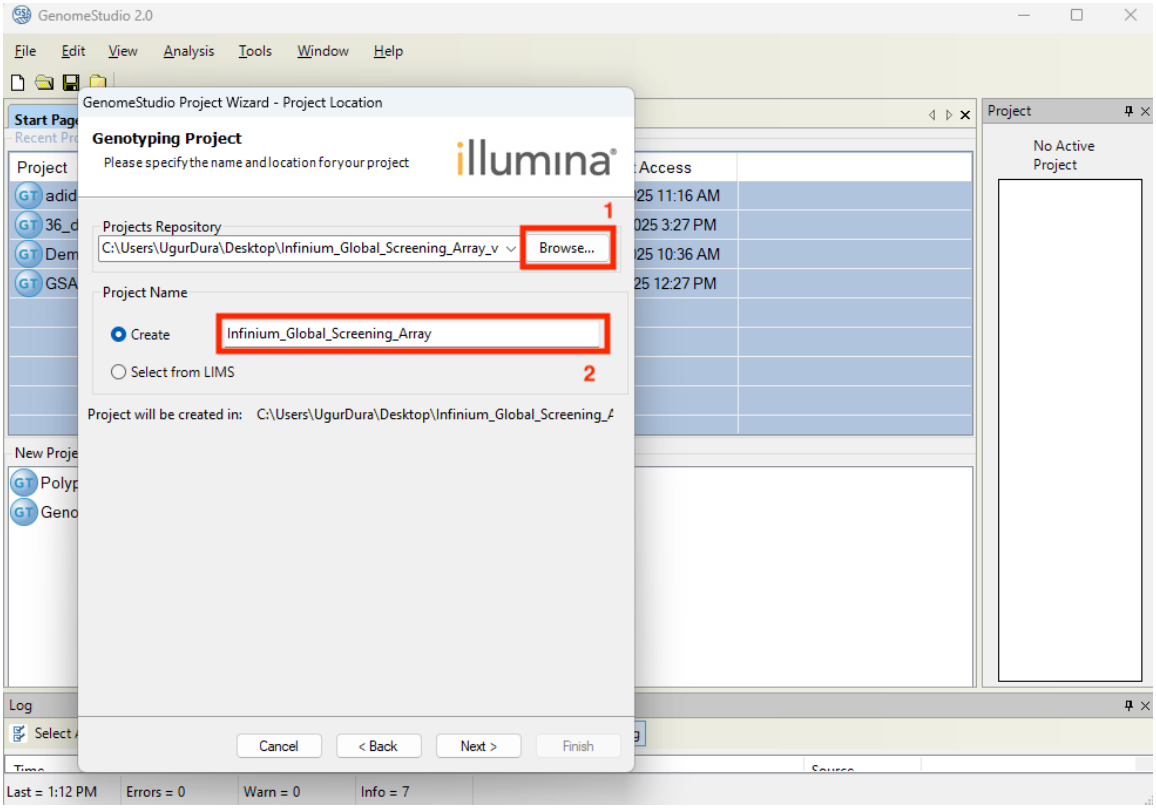

Figure 3.3: Entering the project name and selecting the save location. Use a descriptive name that reflects the dataset (e.g., Infinium_Global_Screening_Array).

3.2.4 Specify Data and Manifest Repositories

After naming your project, the wizard will ask you to specify how to load your samples and where to find the necessary files. This step involves three key inputs: the sample sheet, the data repository (where .idat files are located), and the manifest repository (where .bpm and .egt files are located).

Select Sample Loading Method

First, you need to tell Genome Studio how you want to load your samples.

1. The wizard will display: *"Please specify the samples you want to load by identifying the sample sheet and associated data and manifest repositories"*
2. Select **Use sample sheet to load sample intensities** (see Figure 3.4)
3. Click **Next** to proceed to the repository specification screen

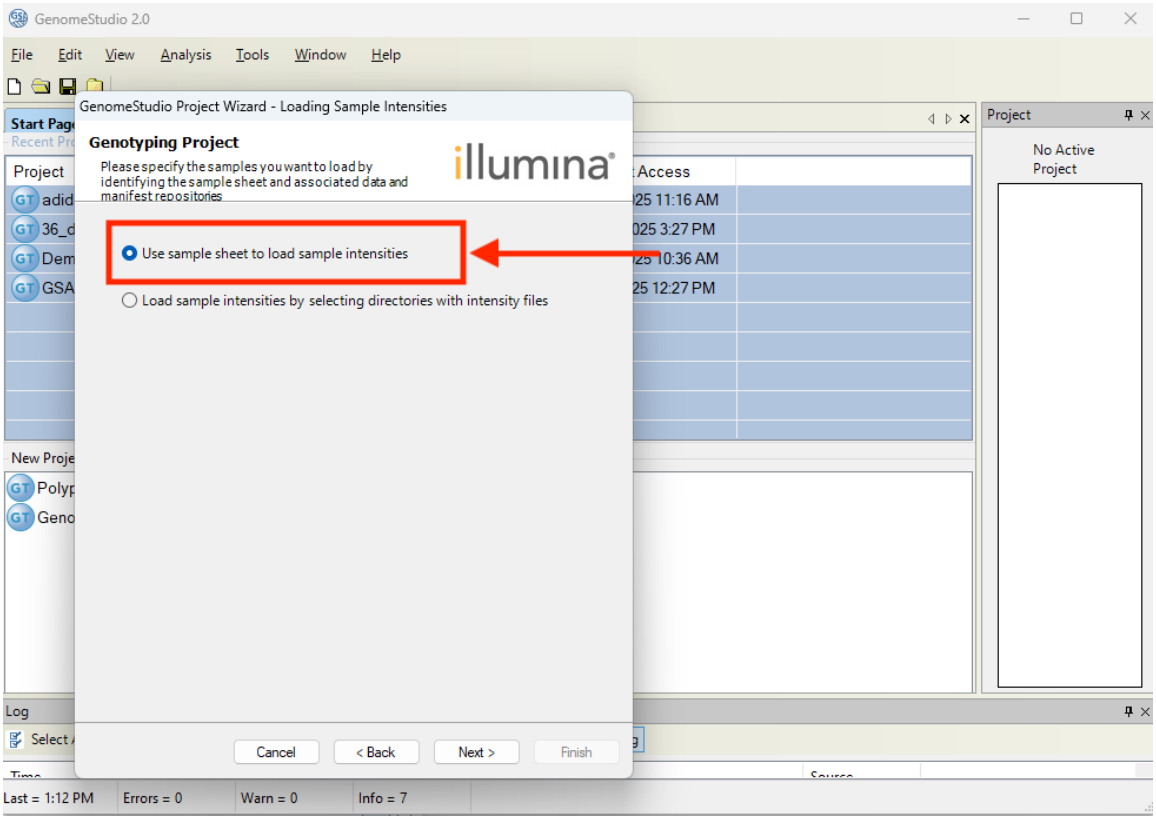

Figure 3.4: Selecting the sample loading method. Choose "Use sample sheet to load sample intensities" to link sample identifiers to their corresponding .idat files using a CSV sample sheet.

Provide Sample Sheet Path

After clicking Next, you will see three input fields for specifying file locations. Start with the sample sheet.

1. In the **Sample Sheet** field, click **Browse** (see Figure 3.5)
2. Navigate to: `Infinium_Global_Screening_Array_v4.0/`
3. Select the file: `GSA-48v4-0_D2_SampleSheet_Demo_36.csv`
4. Click **Open**
5. The path will appear in the Sample Sheet field

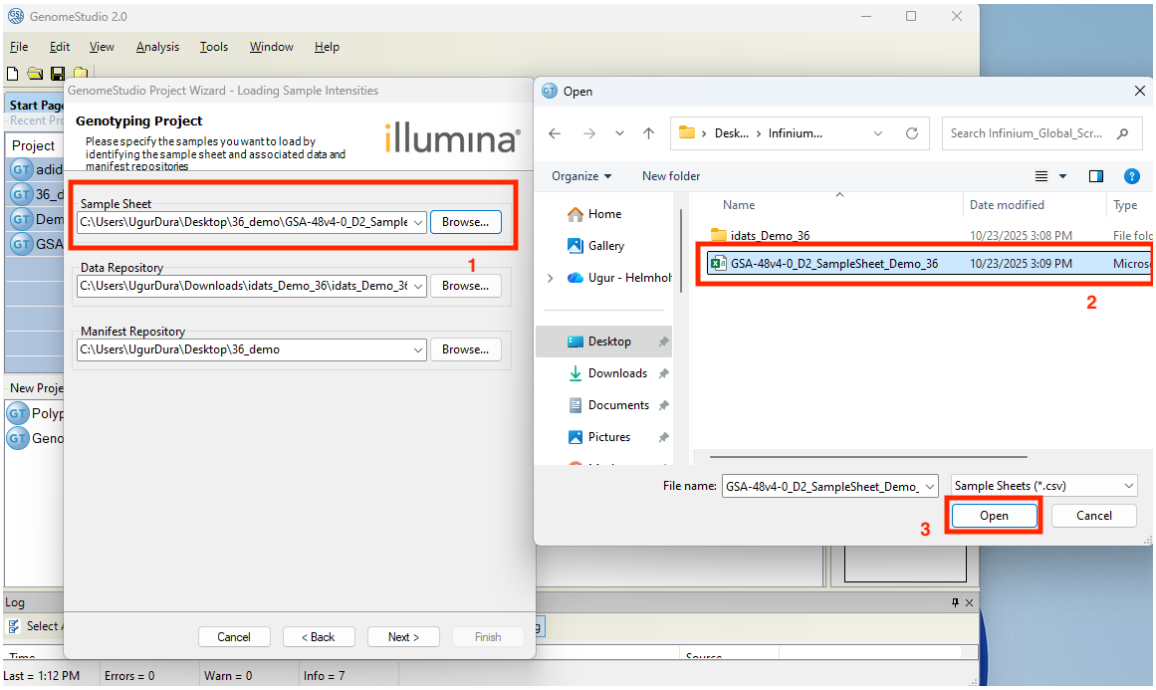

Figure 3.5: Specifying the sample sheet location. Click Browse next to the Sample Sheet field and select the CSV file (GSA-48v4-0_D2_SampleSheet_Demo_36.csv). This file contains the mapping between sample identifiers and their corresponding .idat files.

Provide Data Repository Path

Next, specify where your .idat files (raw intensity data) are located.

1. In the **Data Repository** field, click **Browse** (see Figure 3.6)
2. Navigate to: `Infinium_Global_Screening_Array_v4.0/`
3. Select the folder: `idats_Demo_36/`
4. Click **Select Folder** or **OK**
5. The path will appear in the Data Repository field
6. This folder contains all 72 .idat files (36 samples × 2 channels: Red and Green)

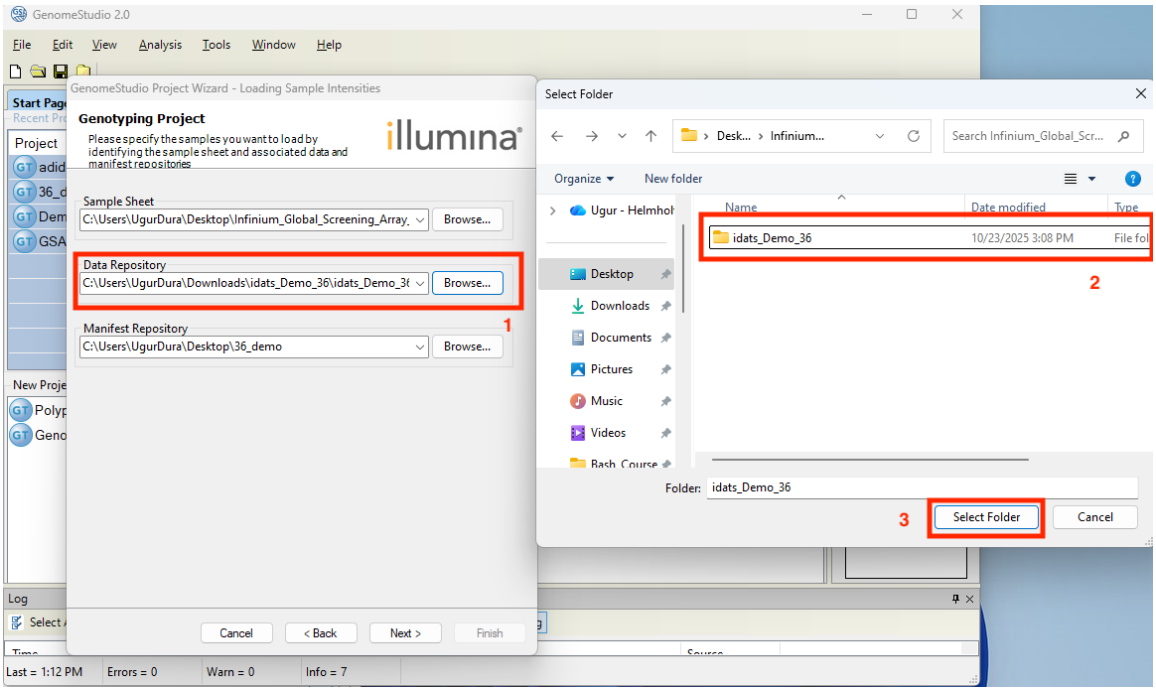

Figure 3.6: Specifying the data repository location. Click Browse next to the Data Repository field and select the folder containing your .idat files (idats_Demo_36/). This folder should contain 72 .idat files: 36 Red channel files and 36 Green channel files.

Provide Manifest Repository Path

Finally, specify where your manifest (.bpm) and cluster (.egt) files are located.

1. In the **Manifest Repository** field, click **Browse** (see Figure 3.7)
2. Navigate to and select the folder: `Infinium_Global_Screening_Array_v4.0/`
3. Click **Select Folder** or **OK**
4. The path will appear in the Manifest Repository field
5. This folder should contain:
   - `GSA-48v4-0_20085471_D2.bpm` (manifest file)
   - `GSA-48v4-0_20085471_D2_ClusterFile.egt` (cluster file)
6. Verify that all three paths (Sample Sheet, Data Repository, Manifest Repository) are correctly specified
7. Click **Next** to proceed to cluster file configuration

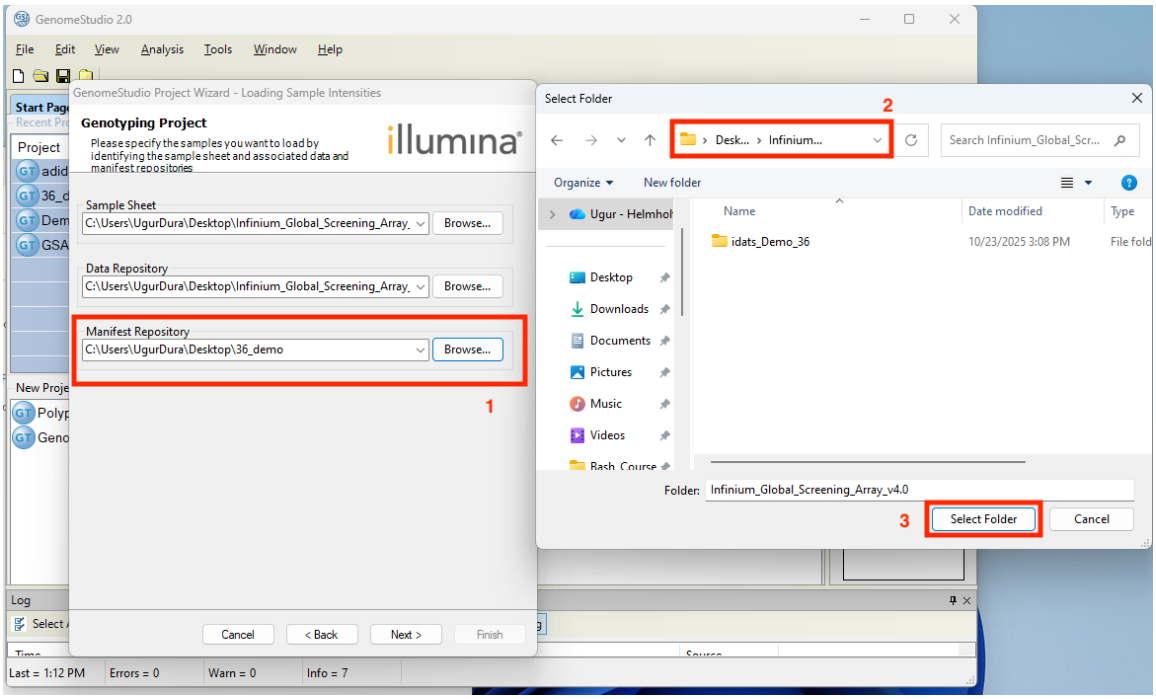

Figure 3.7: Specifying the manifest repository location. Click Browse next to the Manifest Repository field and select the folder (Infinium_Global_Screening_Array_v4.0/) that contains both the manifest file (GSA-48v4-0_20085471_D2.bpm) and cluster file (GSA-48v4-0_20085471_D2_ClusterFile.egt). It is normal not to see .bpm and .egt files in the select folder dialog since it is asking for the parent directory. After verifying all three paths, click Next.

3.2.5 Configure Cluster File and Analysis Settings

After specifying all repository paths, you need to configure the cluster file and analysis settings.

1. On the Cluster File Configuration page, check the box for **Import cluster positions from a cluster file** (see Figure 3.8)
2. Click **Browse** to select the cluster file
3. Navigate to the `Infinium_Global_Screening_Array_v4.0/` folder
4. Select the file: `GSA-48v4-0_20085471_D2_ClusterFile.egt`
5. Click **Open**
6. Verify the following settings:
   - **GenCall Threshold:** 0.15 (default, recommended value)
   - **Calculate Sample and SNP Statistics:** Checked (enabled)
7. Click **Finish** to start the genotype calling process (see Figure 3.9)

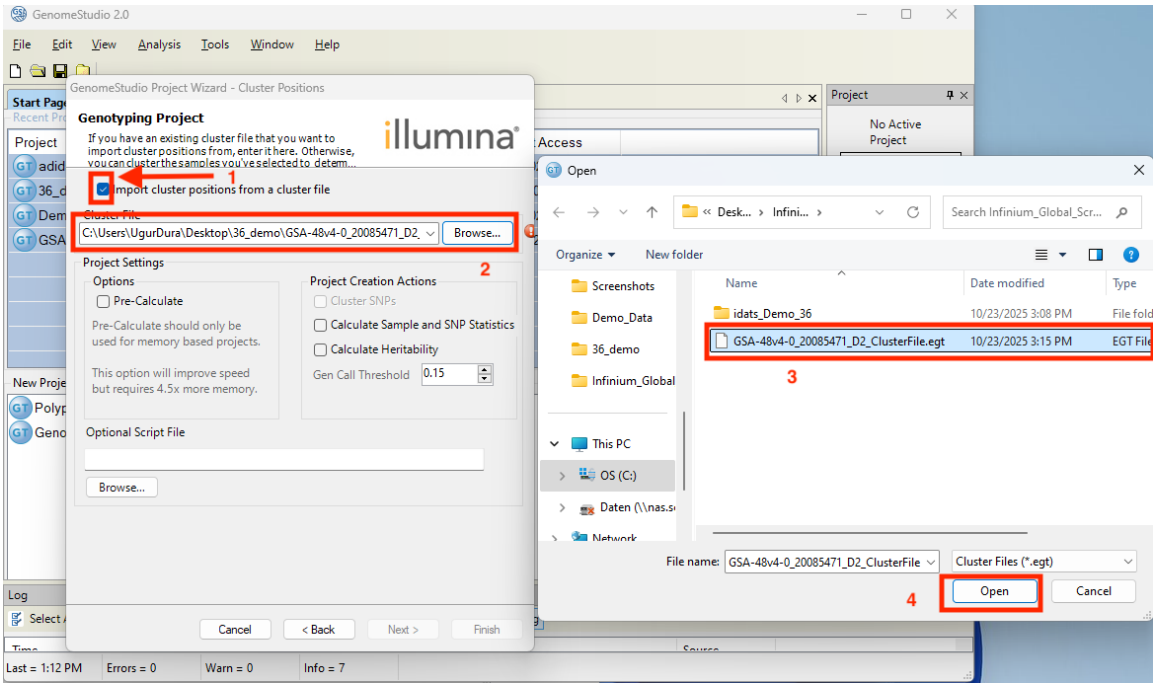

Figure 3.8: Cluster file configuration page. Check "Import cluster positions from a cluster file" and browse to select GSA-48v4-0_20085471_D2_ClusterFile.egt from your working directory. Ensure GenCall Threshold is set to 0.15 and "Calculate Sample and SNP Statistics" is checked.

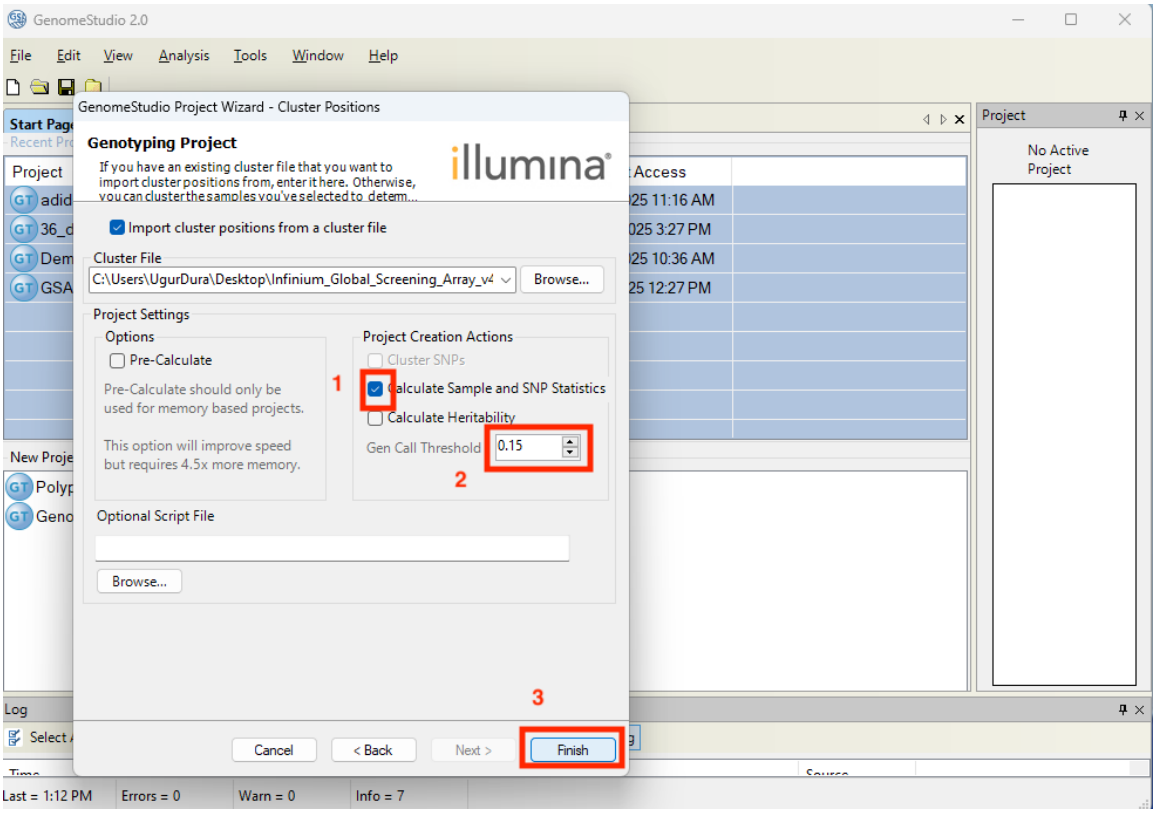

Figure 3.9: Final confirmation page. After verifying all settings including cluster file selection, GenCall threshold (0.15), and statistics calculation option, click Finish to begin the genotype calling process.

Important Settings

GenCall Threshold (0.15): This threshold determines the minimum quality score required for a genotype call to be considered valid. The default value of 0.15 is recommended for most applications. **Calculate Sample and SNP Statistics:** This option must be enabled to generate quality control metrics (call rates, heterozygosity, Hardy-Weinberg equilibrium, etc.) that are essential for data quality assessment.

3.2.6 Genotype Calling Process

After clicking Finish, Genome Studio will begin processing the data automatically.

1. The software will display a loading screen (see Figure 3.10)
2. Genome Studio will:
   - Load all .idat files from the data repository
   - Apply the cluster file for genotype calling
   - Calculate quality metrics (call rates, GC scores, heterozygosity)
   - Generate intensity plots and cluster plots
   - Compute sample and SNP statistics
3. This process may take 10-30 minutes depending on your system specifications
4. A progress bar will show the analysis status
5. Wait for the process to complete before proceeding to the next step

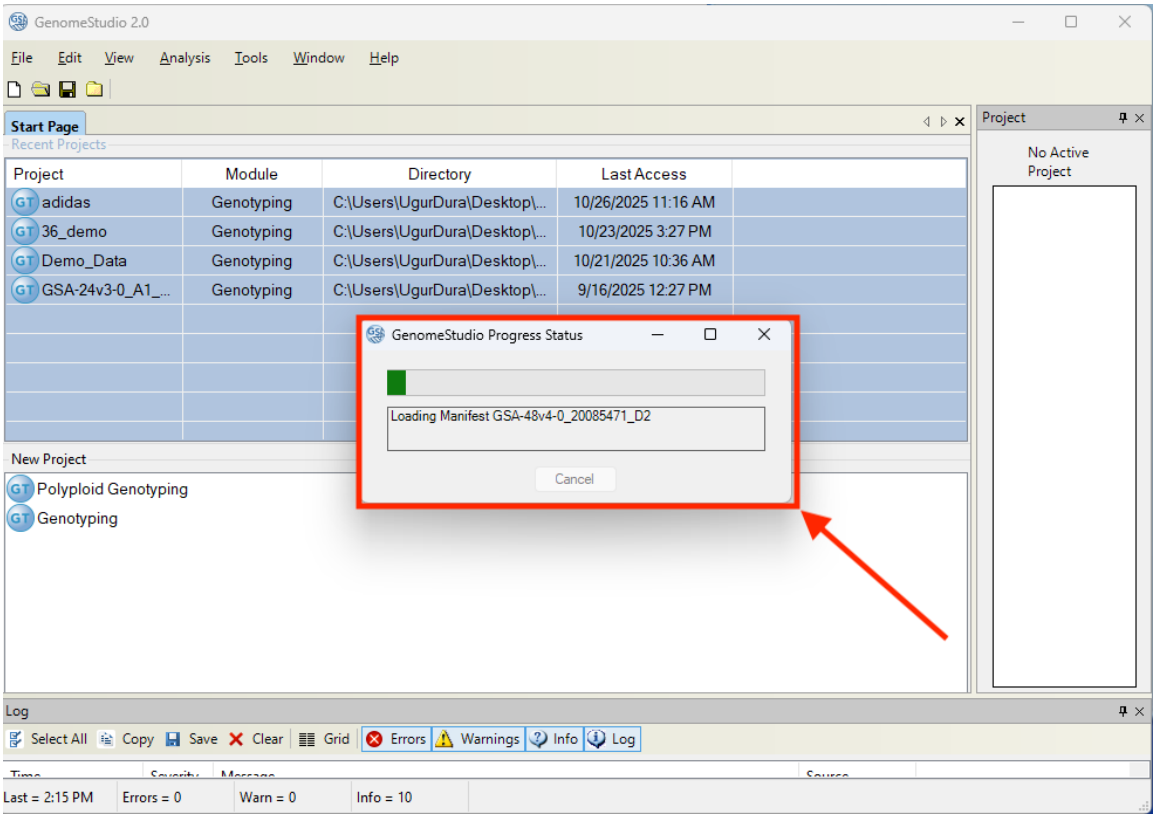

Figure 3.10: Genome Studio loading screen during data processing. The software is reading .idat files, applying the cluster file, and performing genotype calling. This may take 10-30 minutes for 36 samples.

Once the genotype calling process is complete, Genome Studio will prompt you to update quality metrics.

3.2.7 Update Heritability and Reproducibility Errors

After the initial data processing completes, Genome Studio will display a dialog asking about updating quality metrics.

1. A dialog will appear with the message: *"Do you wish to update all heritability and reproducibility errors? This operation may take some time."* (see Figure 3.11)
2. Click **Yes** to update these quality metrics
3. A progress bar will appear showing the calculation status (see Figure 3.12)
4. This process typically takes 5-10 minutes for 36 samples
5. Wait for the process to complete before proceeding

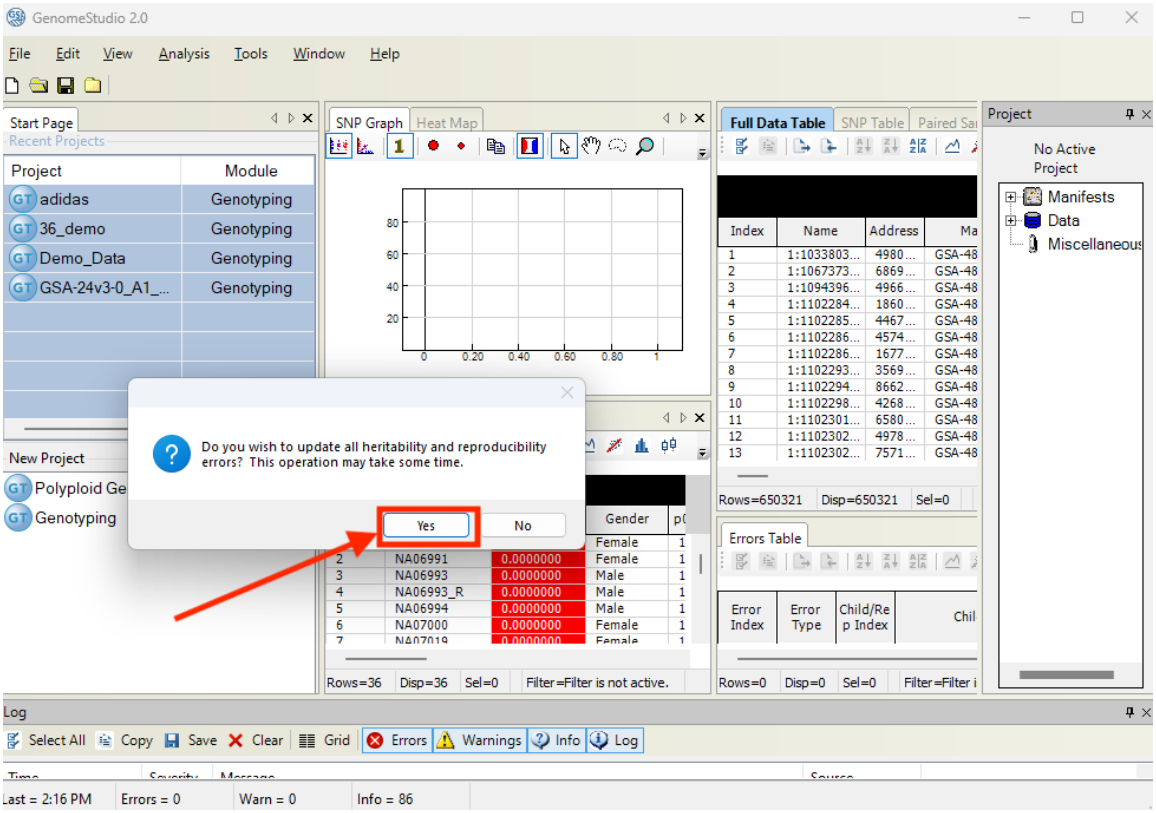

Figure 3.11: Heritability and reproducibility error update prompt. Click Yes to calculate these quality metrics, which are important for assessing data reliability and identifying potential technical issues.

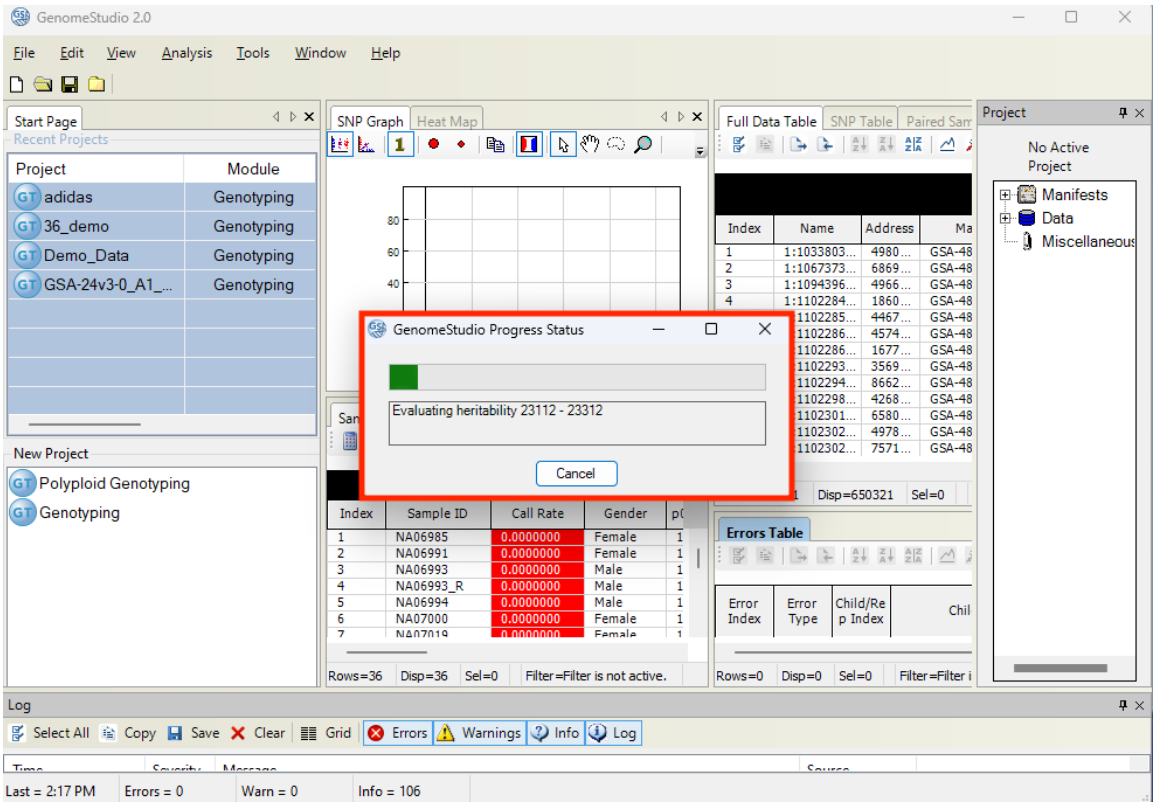

Figure 3.12: Progress bar showing the calculation of heritability and reproducibility errors. This process analyzes replicate samples and family relationships to compute quality metrics. Wait for completion before proceeding.

About Heritability and Reproducibility Errors

Heritability Errors: Measure the consistency of genotype calls across related samples (parent-offspring, siblings). High heritability errors may indicate genotyping quality issues.
Reproducibility Errors: Measure the consistency of genotype calls across technical replicates of the same sample. Low reproducibility indicates poor technical quality.
These metrics are essential for quality control and should always be calculated.

Once the heritability and reproducibility error calculations are complete, Genome Studio will display the main genotyping interface with multiple data views (see Figure 3.13).

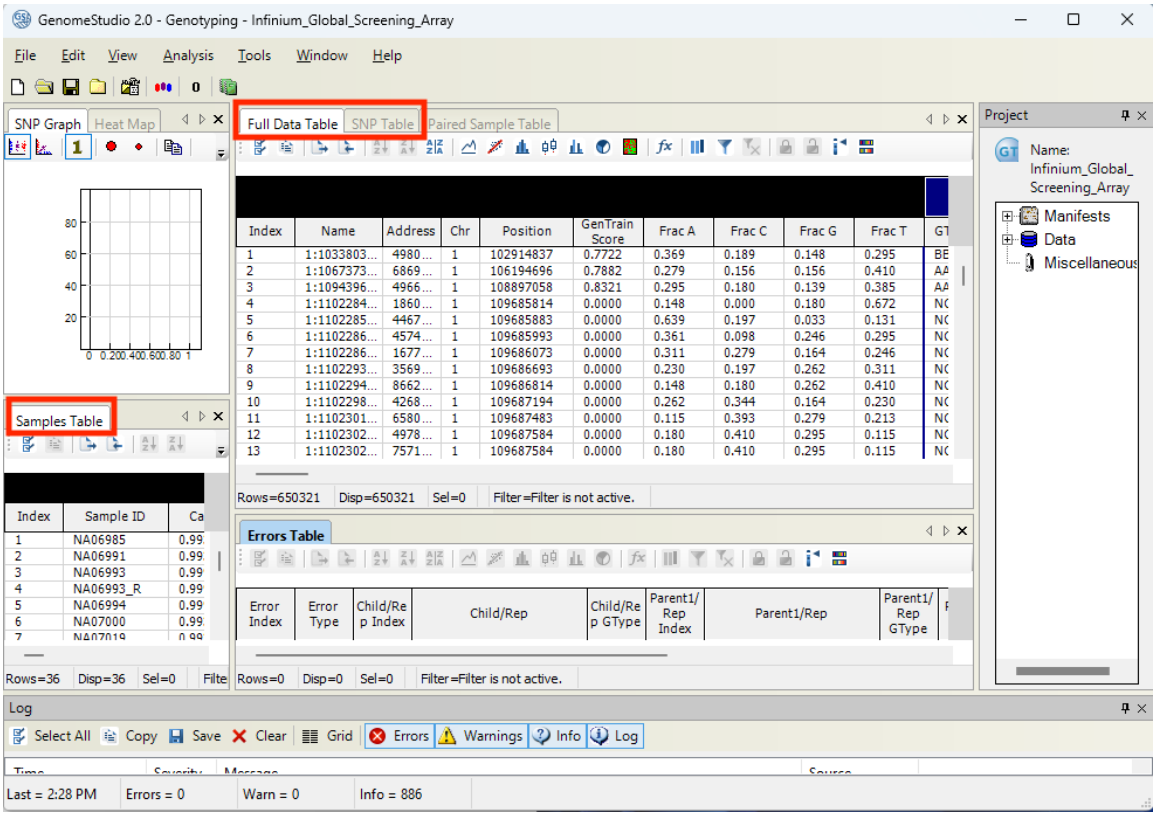

Figure 3.13: Genome Studio genotyping module interface after successful project loading and quality metrics calculation. The interface shows multiple tabs including Full Data Table, SNP Table, and Samples Table, which contain all the genotyping results and quality metrics.

3.3 Step 2: Full Data Table Column Configuration

After the genotype calling is complete, you need to configure the columns in the Full Data Table to ensure all required data fields are included for export.

3.3.1 Access Full Data Table View

1. Once genotype calling is complete, navigate to the **Full Data Table** tab
2. This table displays all genotyping data in a spreadsheet format
3. By default, not all columns are visible

3.3.2 Configure Required Columns

1. Right-click on the column header area (see Figure 3.14)
2. Select **Choose Columns** or **Column Chooser**
3. A dialog box will appear with available columns on the right (Hidden Columns) and selected columns on the left (Shown Columns)

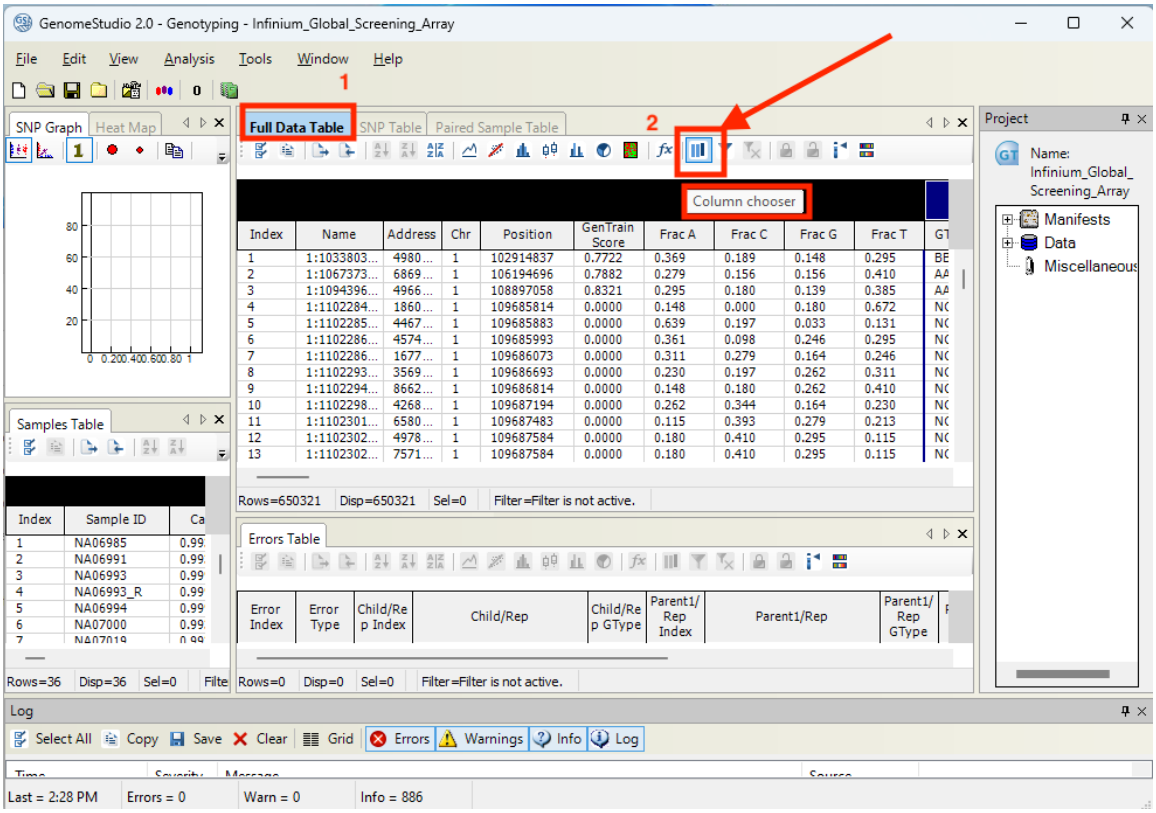

Figure 3.14: Opening the Column Chooser dialog. This allows you to select which data fields to include in the Full Data Table export.

3.3.3 Select Essential Columns

Ensure the following columns are selected for the Digital Karyotyping Pipeline. The most critical columns are **B Allele Freq** and **Log R Ratio**:

Column Name	Description
Index	Row index number
Name	Marker identifier (rsID or probe name)
Address	Probe address on the array
Chr	Chromosome
Position	Genomic position (bp)
GenTrain Score	Cluster quality score
Frac A	Fraction of A allele
Frac C	Fraction of C allele
Frac G	Fraction of G allele
Frac T	Fraction of T allele
GType	Genotype call
Score	Genotype quality score
Theta	Normalized angle (allelic ratio)
R	Normalized intensity (total signal)
B Allele Freq	B Allele Frequency (BAF) [10]
Log R Ratio	Log R Ratio (LRR) [10]

Table 3.1: Essential columns required for the Full Data Table export

Adding B Allele Freq Column

1. In the Column Chooser dialog, locate **B Allele Freq** in the right panel (Hidden Columns)
2. Select **B Allele Freq** (see Figure 3.15)
3. Click the **Show** button to move it to the left panel (Shown Columns)

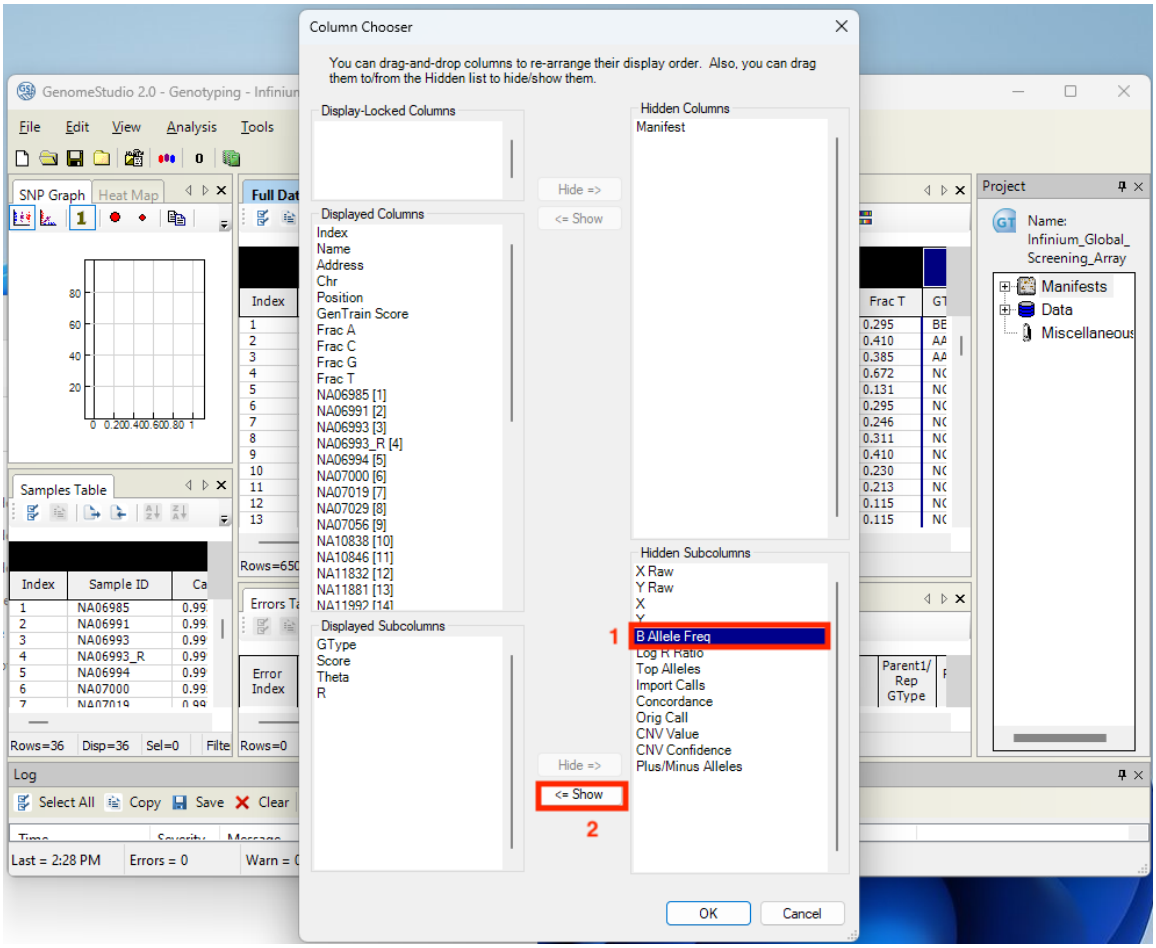

Figure 3.15: Selecting B Allele Freq from the Hidden Columns panel. Click the Show button to add it to the visible columns in the Full Data Table.

Adding Log R Ratio Column

1. In the Column Chooser dialog, locate **Log R Ratio** in the right panel (Hidden Columns)
2. Select **Log R Ratio** (see Figure 3.16)
3. Click the **Show** button to move it to the left panel (Shown Columns)
4. Click **OK** to apply the column configuration

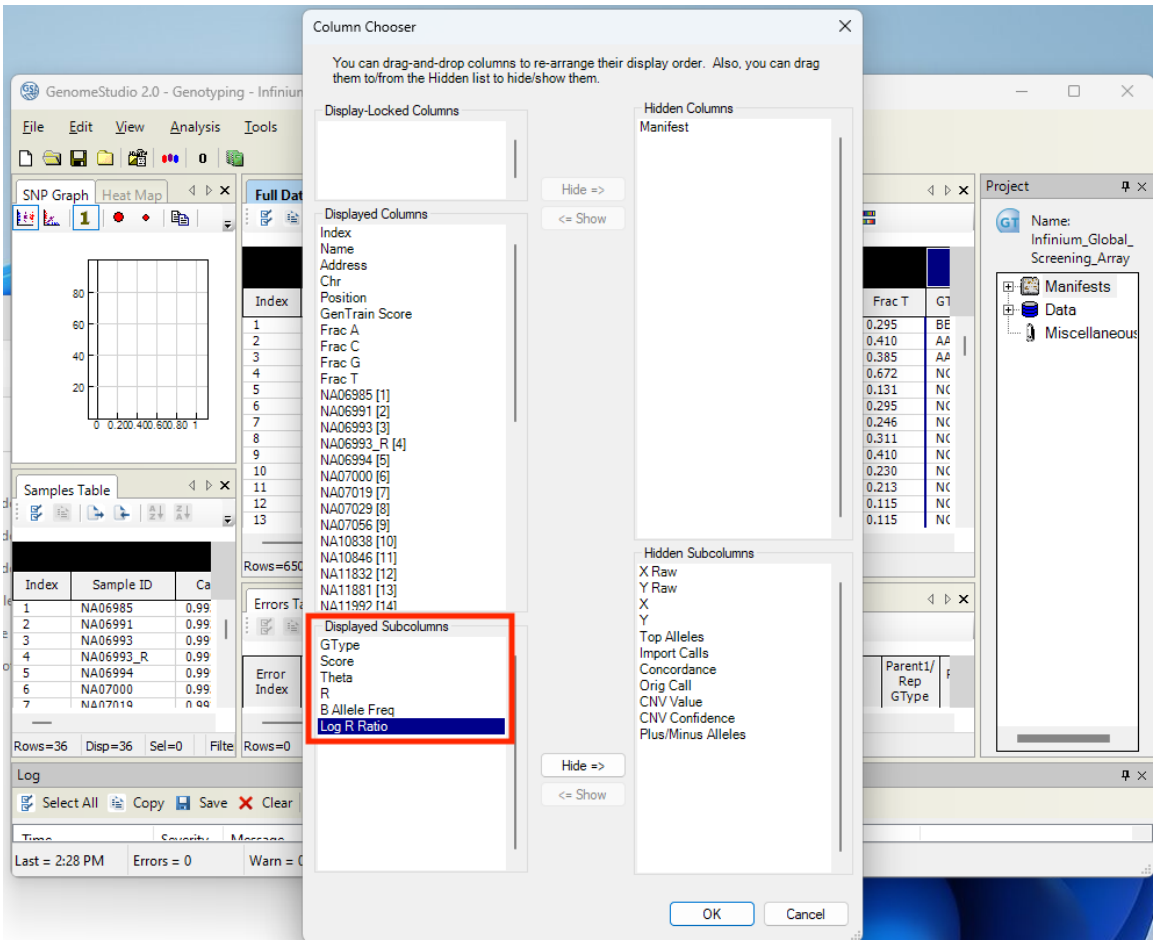

Figure 3.16: Selecting Log R Ratio from the Hidden Columns panel. After clicking Show, click OK to apply all column changes to the Full Data Table.

Critical Step

Important: Ensure all columns in Table 3.1 are selected before exporting. Missing columns will cause errors in the Digital Karyotyping Pipeline. Pay special attention to:

- B Allele Freq (BAF) - Required for CNV detection
- Log R Ratio (LRR) - Required for copy number analysis
- Chr and Position - Required for genomic mapping

3.4 Step 3: Exporting Data Tables

Now that the project is created and columns are configured, you can export the required data tables for the Digital Karyotyping Pipeline.

3.4.1 Export Full Data Table

After configuring all required columns, you can now export the Full Data Table.

Create Data Directory

Before exporting, create a `data/` directory within your project folder to organize all exported files:

```
Infinium_Global_Screening_Array_v4.0/
|-- GSA-48v4-0_20085471_D2.bpm
|-- GSA-48v4-0_20085471_D2_ClusterFile.egt
|-- GSA-48v4-0_D2_SampleSheet_Demo_36.csv
|-- idats_Demo_36/
`-- data/                          <- Create this directory
    `-- Full_Data_Table.txt        <- Export files here
```

Export Process

1. In the Full Data Table view, click the **Export Data** button in the toolbar (see Figure 3.17)
2. A file save dialog will appear
3. Navigate to the `data/` directory you created under your project folder
4. Save as: `Full_Data_Table.txt`
5. Click **Save**

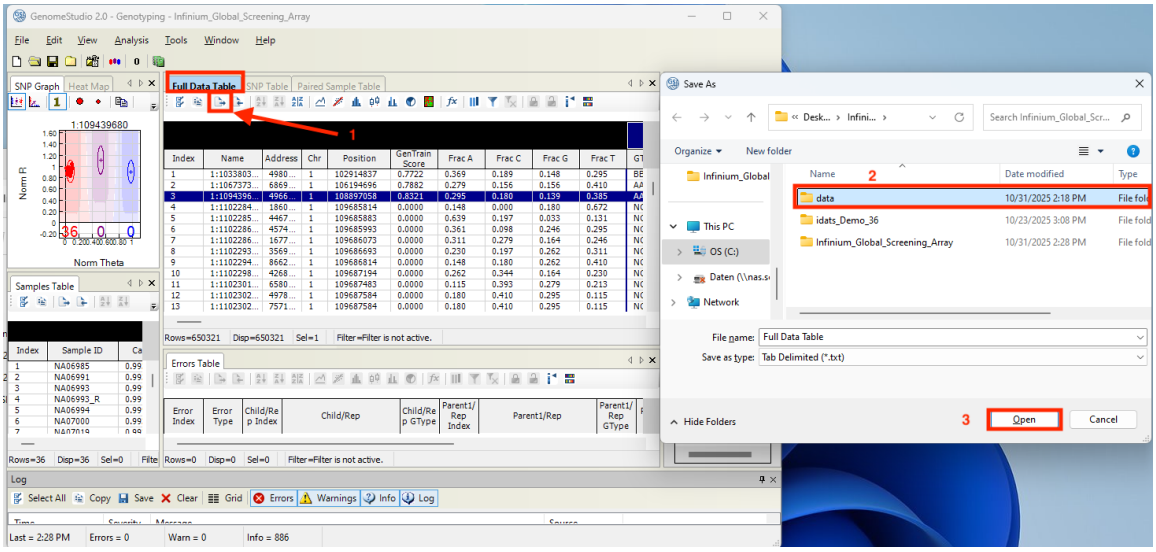

Figure 3.17: Clicking the Export Data button in the Full Data Table view and selecting the save location. Navigate to the `data/` directory within your project folder and save the file as `Full_Data_Table.txt`.

Export Confirmation

1. After clicking Save, a dialog will appear asking: *"Currently, only the selected rows and columns will be exported. Would you prefer to Export the entire table?"* (see Figure 3.18)
2. Click **Yes** to export the entire table (all samples and all SNPs)
3. The export process will begin (see Figure 3.19 for progress)

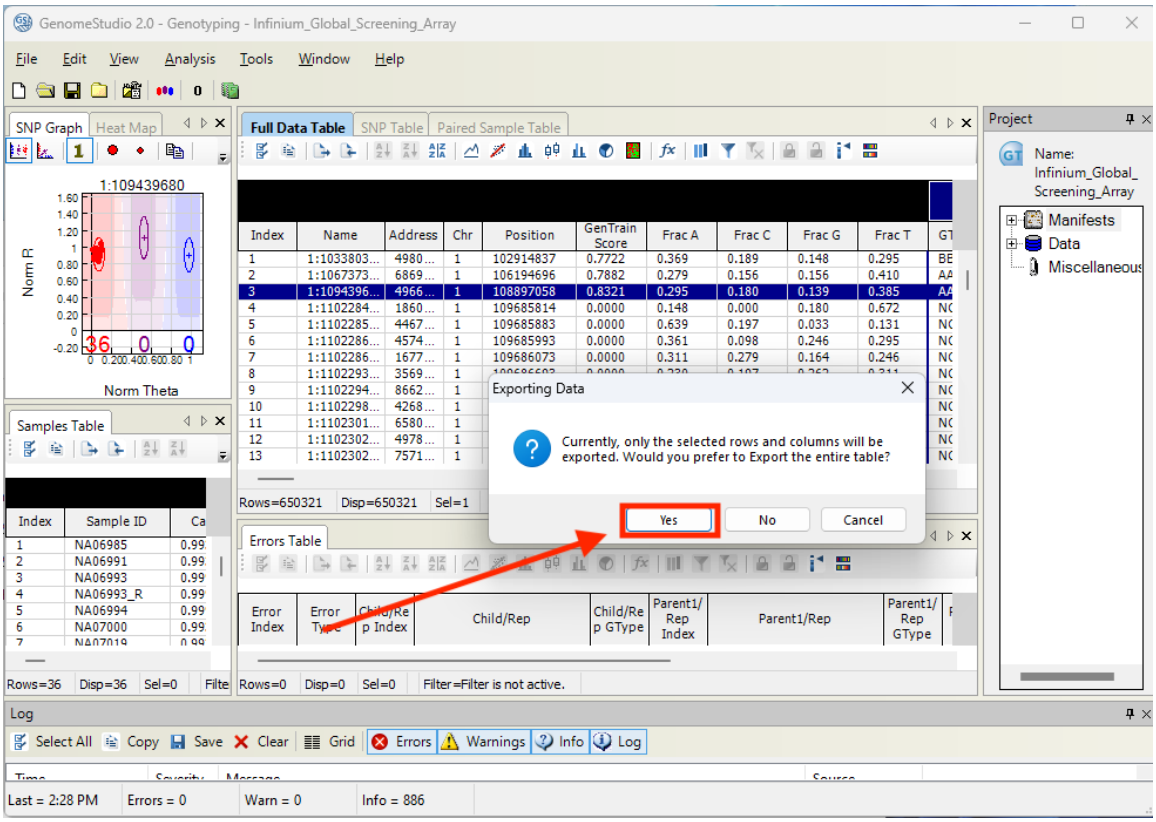

Figure 3.18: Export confirmation dialog. Click Yes to export the entire Full Data Table including all samples and SNPs with all configured columns.

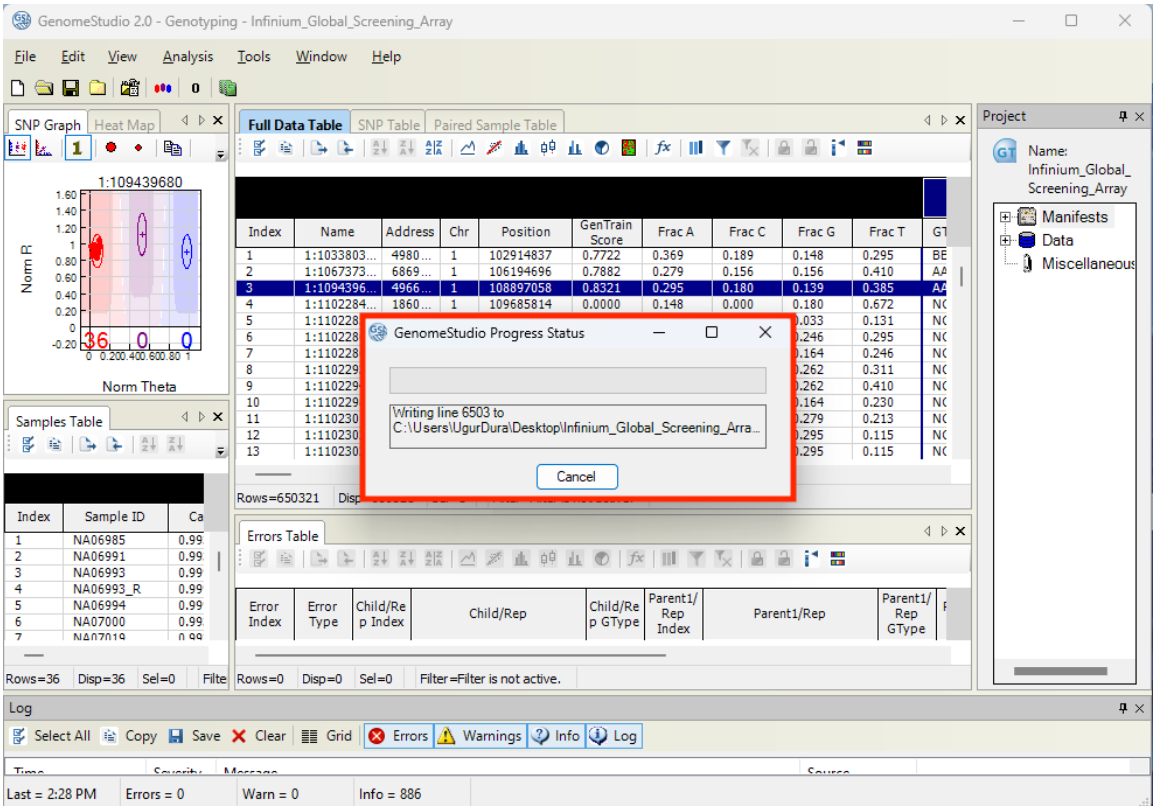

Figure 3.19: Full Data Table export progress bar. The export may take several minutes depending on the dataset size (36 samples × 700,000 SNPs for this demo).

Export Time Estimate

The Full Data Table export typically takes:

- **Demo dataset (36 samples):** 5-10 minutes
- **Larger datasets (100+ samples):** 15-30 minutes
- File size: Approximately 1-2 GB for this demo dataset

Do not close Genome Studio during the export process.

3.4.2 Export SNP Table

After exporting the Full Data Table, you need to export the SNP Table which contains marker-level statistics.

1. Navigate to the **SNP Table** tab in Genome Studio
2. Click the **Export Data** button in the toolbar (see Figure 3.20)
3. A file save dialog will appear
4. Navigate to the `data/` directory (same location as Full Data Table)
5. Save as: `SNP_Table.txt`
6. Click **Save**
7. A dialog will appear asking: *"Would you prefer to Export the entire table?"*
8. Click **Yes** to export all SNPs with all statistics
9. The export process will begin (see Figure 3.21 for export progress)

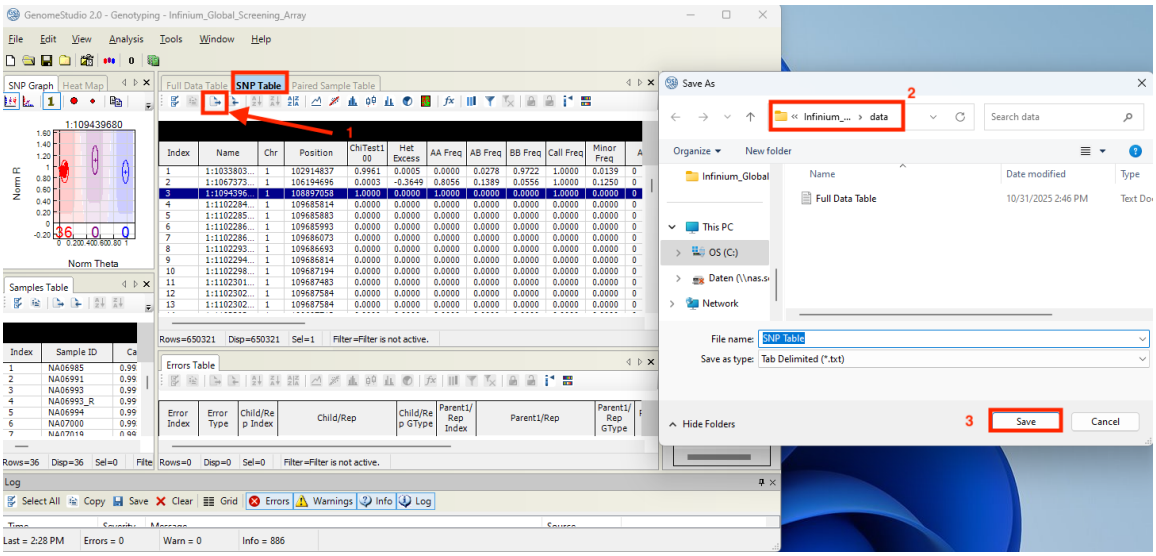

Figure 3.20: Exporting the SNP Table. Click the Export Data button in the SNP Table tab and save to the `data/` directory. This table contains marker-level statistics including call rates, allele frequencies, and Hardy-Weinberg equilibrium p-values.

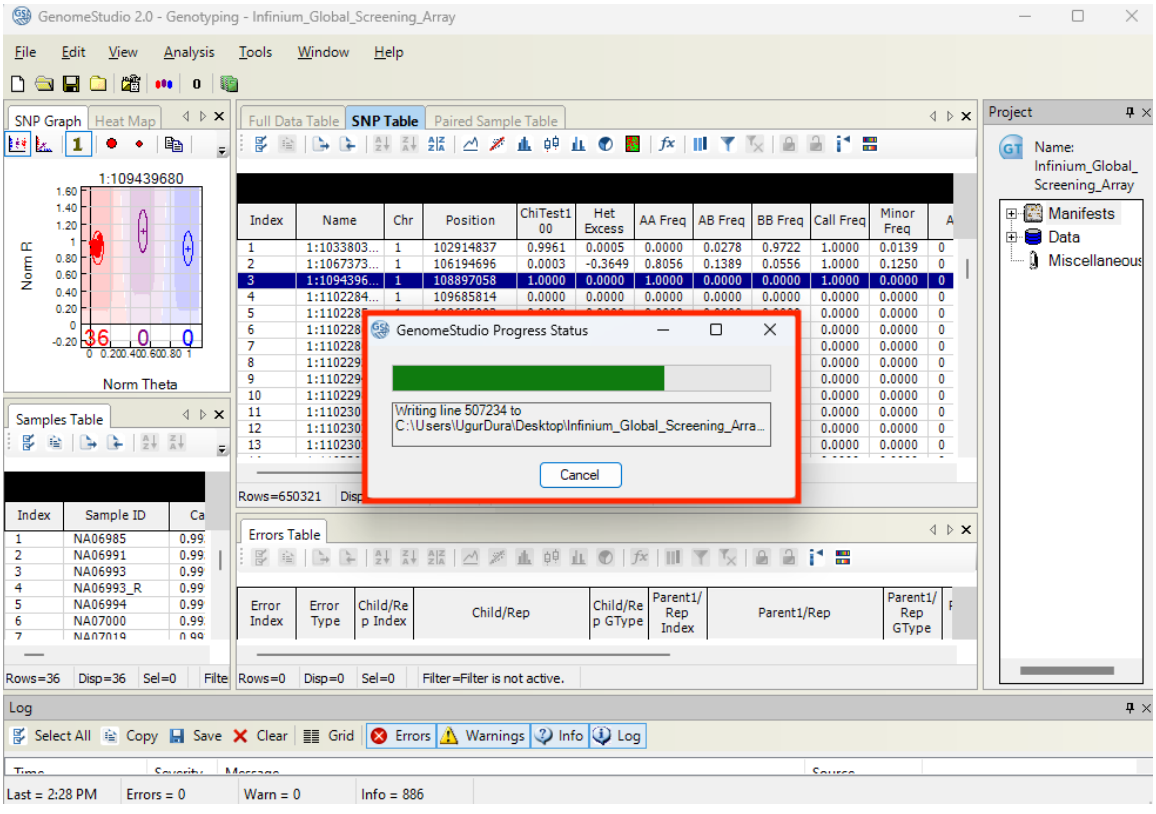

Figure 3.21: Export progress indicator for SNP Table. The export typically takes 2-5 minutes for 700,000 SNPs.

3.4.3 Export Samples Table

Finally, export the Samples Table which contains sample-level quality control metrics.

1. Navigate to the **Samples Table** tab in Genome Studio
2. Click the **Export Data** button in the toolbar (see Figure 3.22)
3. A file save dialog will appear
4. Navigate to the `data/` directory (same location as previous exports)
5. Save as: `Samples_Table.txt`
6. Click **Save**
7. A dialog will appear asking: *"Would you prefer to Export the entire table?"*
8. Click **Yes** to export all samples with all QC metrics
9. The export completes quickly (typically under 1 minute for 36 samples)

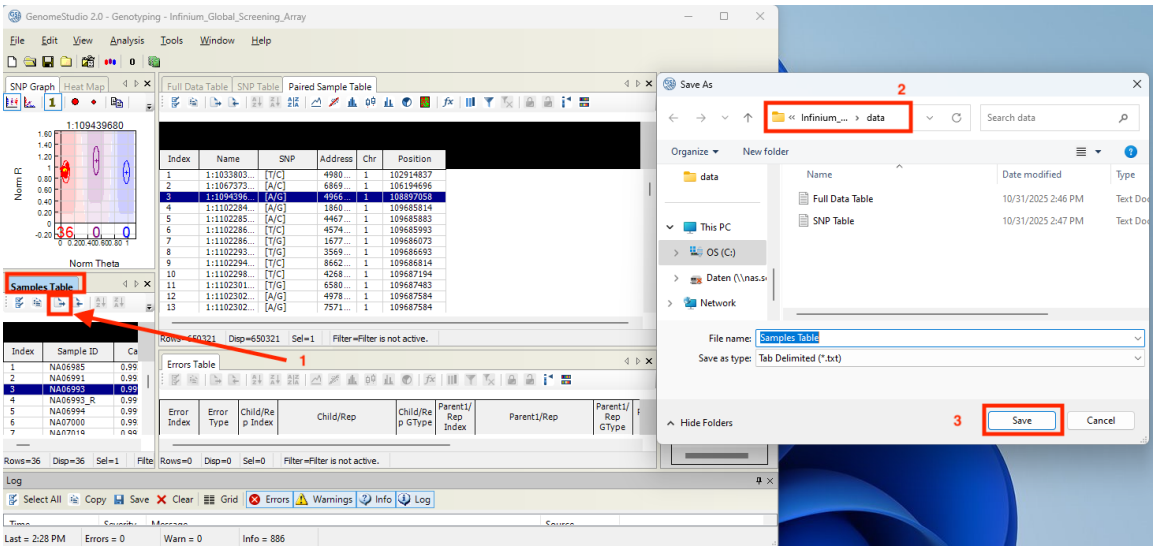

Figure 3.22: Exporting the Samples Table. Click the Export Data button in the Samples Table tab and save to the `data/` directory. This table contains sample-level QC metrics (call rates, heterozygosity, etc.) essential for quality control and filtering.

Data Export Complete

After exporting all three tables, your `data/` directory should contain:

- `Full_Data_Table.txt` - Complete genotyping data with LRR and BAF
- `SNP_Table.txt` - Marker-level statistics and QC metrics
- `Samples_Table.txt` - Sample-level QC metrics

These three files are required inputs for the Digital Karyotyping Pipeline.

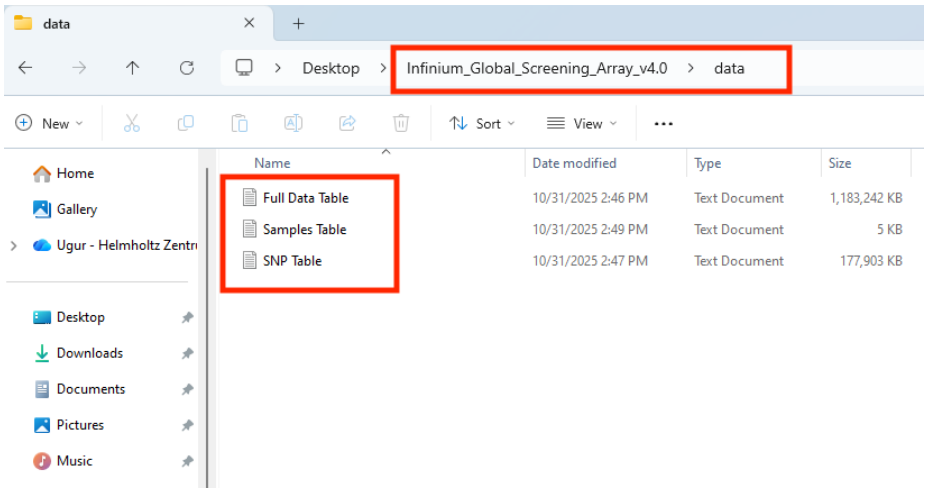

Figure 3.23: Contents of the `data/` directory after exporting all three tables (Full Data Table, SNP Table, and Samples Table). These are the core Genome Studio exports required for the pipeline.

Export Tips

Best Practices for Data Export:

- Create a dedicated `./data/` folder for all output files
- Verify file sizes after export (Full Data Table should be largest)

3.4.4 Export Process

3.5 Step 4: Generating PLINK Files

The final step is to generate PLINK format files using the PLINK Input Report Plug-in v2.1.4 installed in Chapter 1. This plug-in converts Genome Studio genotyping data into PLINK-compatible formats (.ped, .map, .bed, .bim, .fam) required by the Digital Karyotyping Pipeline.

3.5.1 Access Report Wizard

1. In Genome Studio, go to **Analysis** → **Reports** → **Report Wizard** (see Figure 3.24)
2. The Report Wizard window will open with several report options

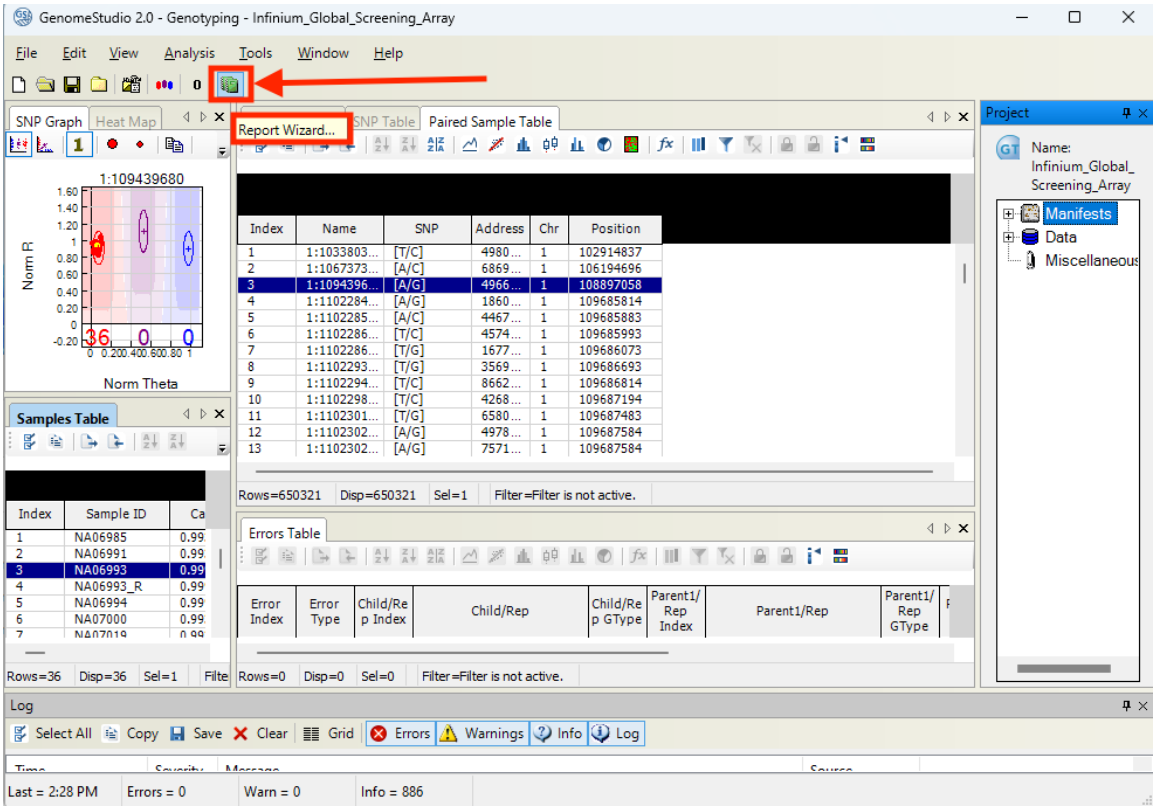

Figure 3.24: Accessing the Report Wizard. Click on Analysis → Reports → Report Wizard to open the report generation interface.

3.5.2 Select PLINK Input Report

1. In the Report Wizard window, you will see several report options (see Figure 3.25)
2. Under the **Custom Reports** section, select **PLINK Input Report 2.1.4**
   - If you do not see PLINK in the list, verify that the PLINK Input Report Plug-in was installed correctly (see Chapter 1)
3. Click **Next** to proceed

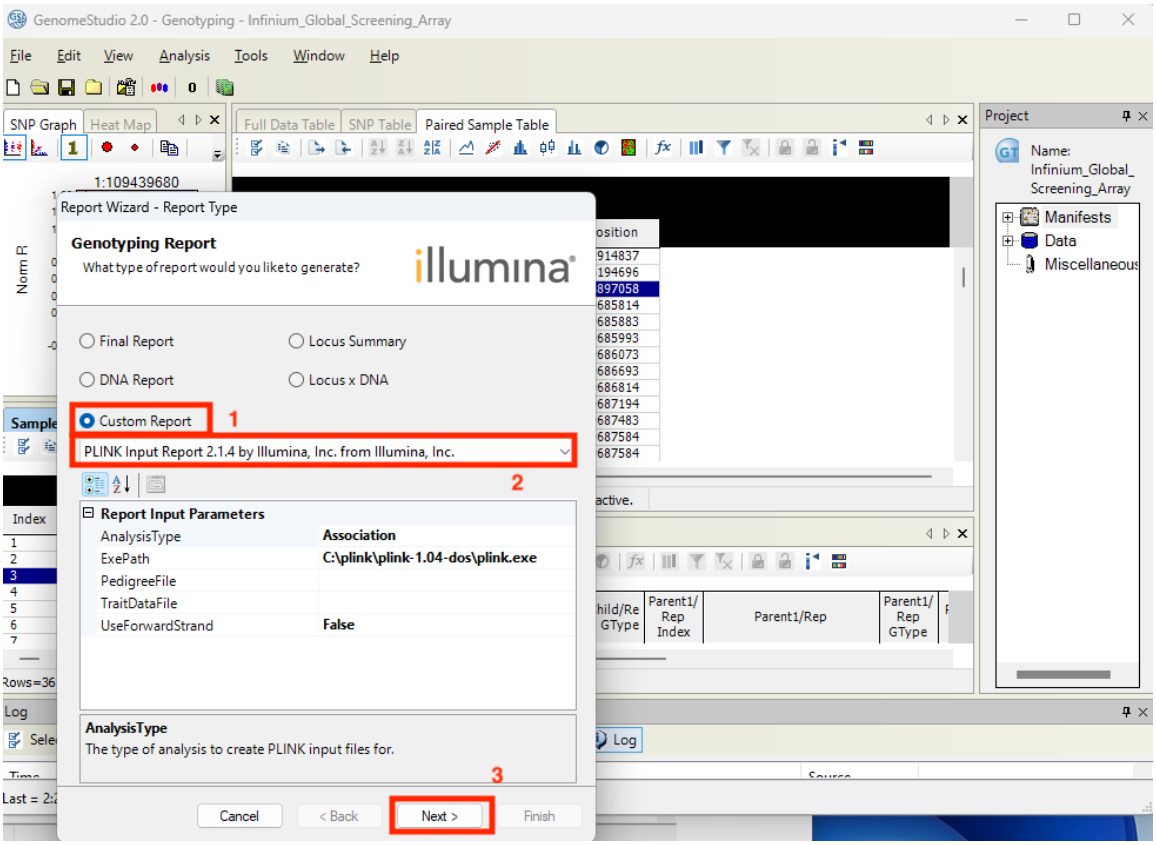

Figure 3.25: Selecting PLINK Input Report from the Custom Reports section. The version shown should be 2.1.4 if the plug-in was installed correctly.

3.5.3 Select Samples to Include

1. The wizard will ask: *"Which samples would you like to include in your report?"* (see Figure 3.26)
2. Select **All samples**
3. Click **Next** to proceed

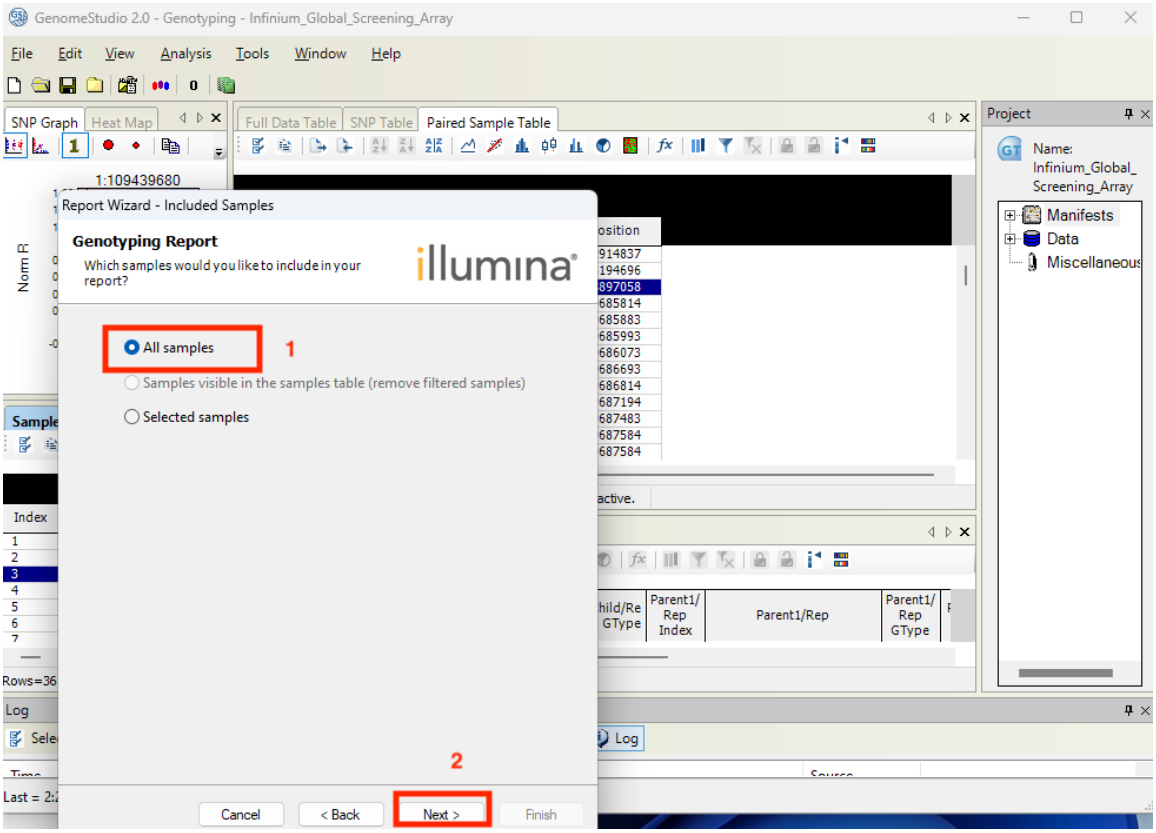

Figure 3.26: Selecting samples for PLINK export. Choose "All samples" to include all 36 demo samples in the output.

3.5.4 Select Sample Groups

1. The wizard will ask: *"This project has various sample groups. Please select the ones you want to include in the report."* (see Figure 3.27)
2. Select **all available options**:
   - CEU
   - CEU_Rep
   - CEU_Unrel
3. Click **Next** to proceed

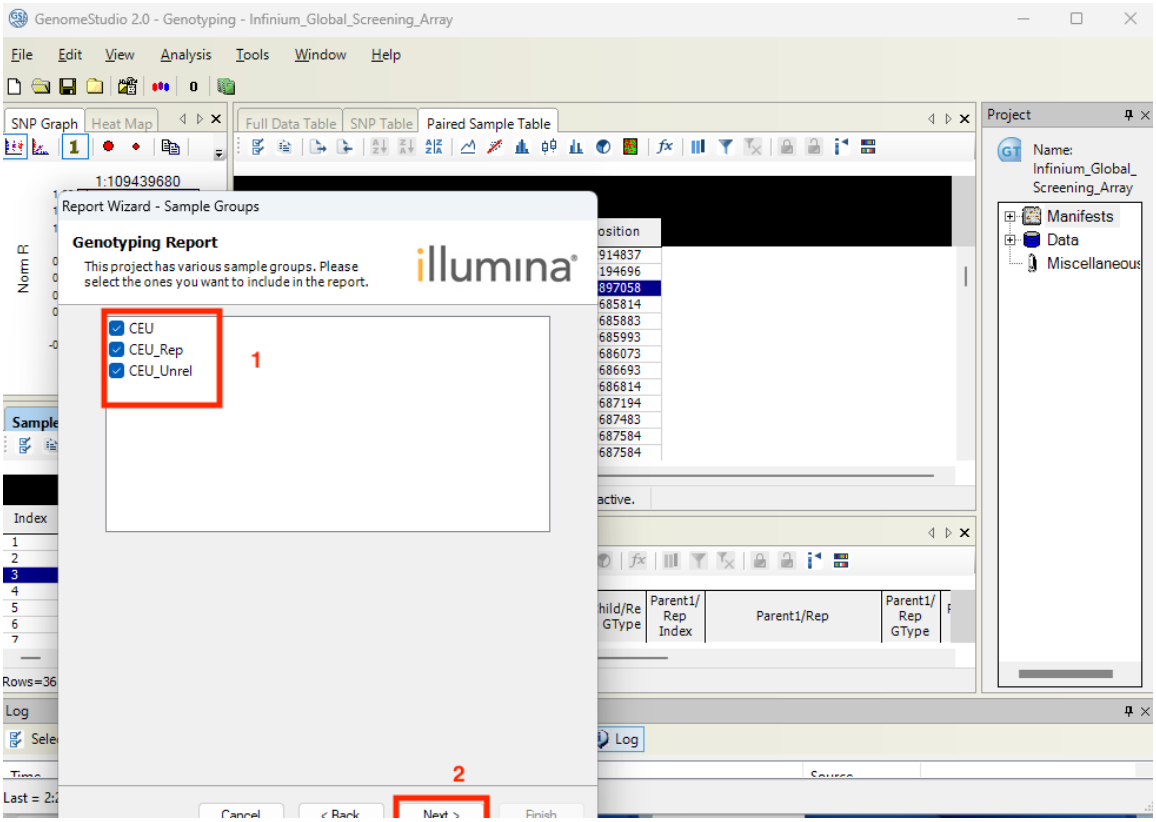

Figure 3.27: Selecting sample groups for PLINK export. Choose all available groups (CEU, CEU_Rep, and CEU_Unrel) to include all samples.

3.5.5 Select SNPs to Include

1. The wizard will ask: *"Which SNPs would you like to include in your report?"* (see Figure 3.28)
2. Select **Include zeroed SNPs in the report**
   - This ensures all SNPs are included, even those with zero intensity values
3. Click **Next** to proceed

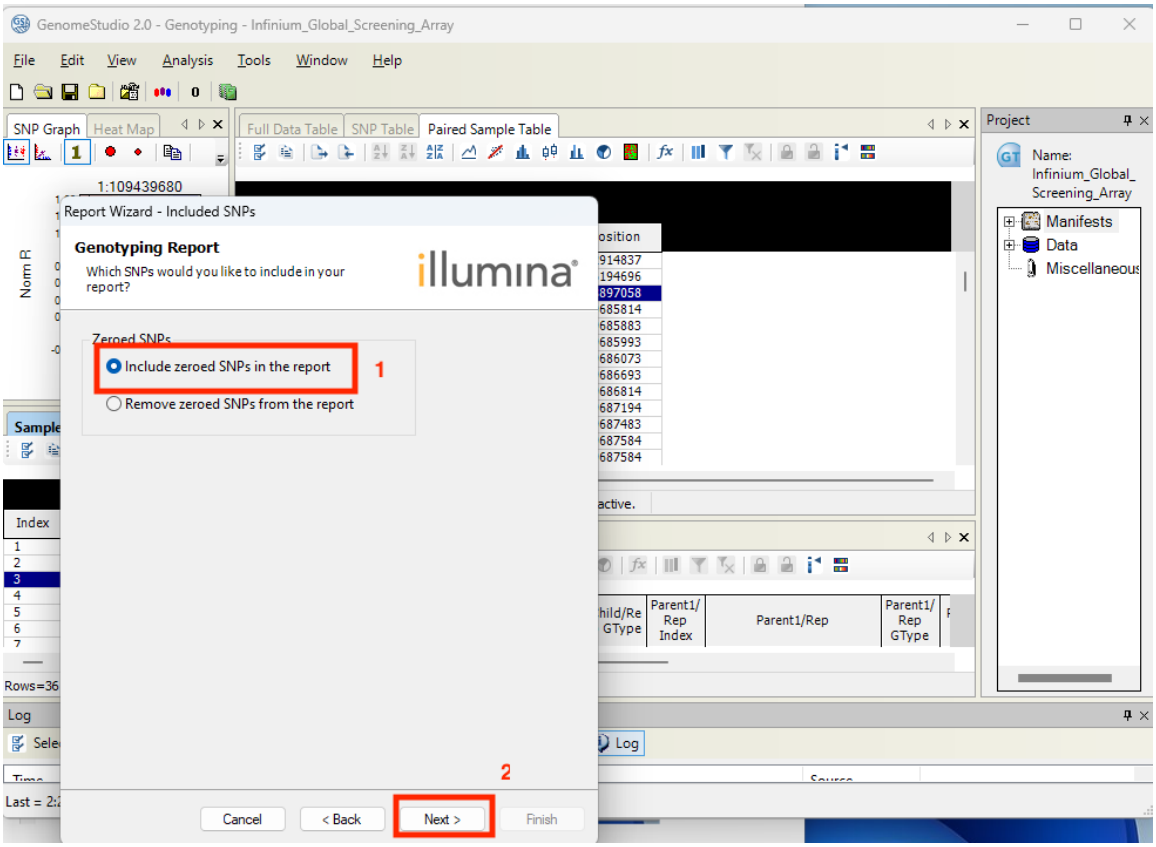

Figure 3.28: Selecting SNPs for PLINK export. Choose "Include zeroed SNPs in the report" to ensure all SNPs are exported, including those with zero intensity values.

3.5.6 Configure Output Path and Report Name

1. The wizard will ask for the output path and report name (see Figure 3.29)
2. **Report Name**: Use the default name `Infinium_Global_Screening_Array_Custom`
   - The report file itself is not needed for the Digital Karyotyping Pipeline, only the PLINK output files
3. **Output Path**: Click **Browse** and navigate to the `data/` directory you created earlier
   - This is the same directory where you saved the Full Data Table, SNP Table, and Samples Table
4. Click **Finish** to start PLINK file generation

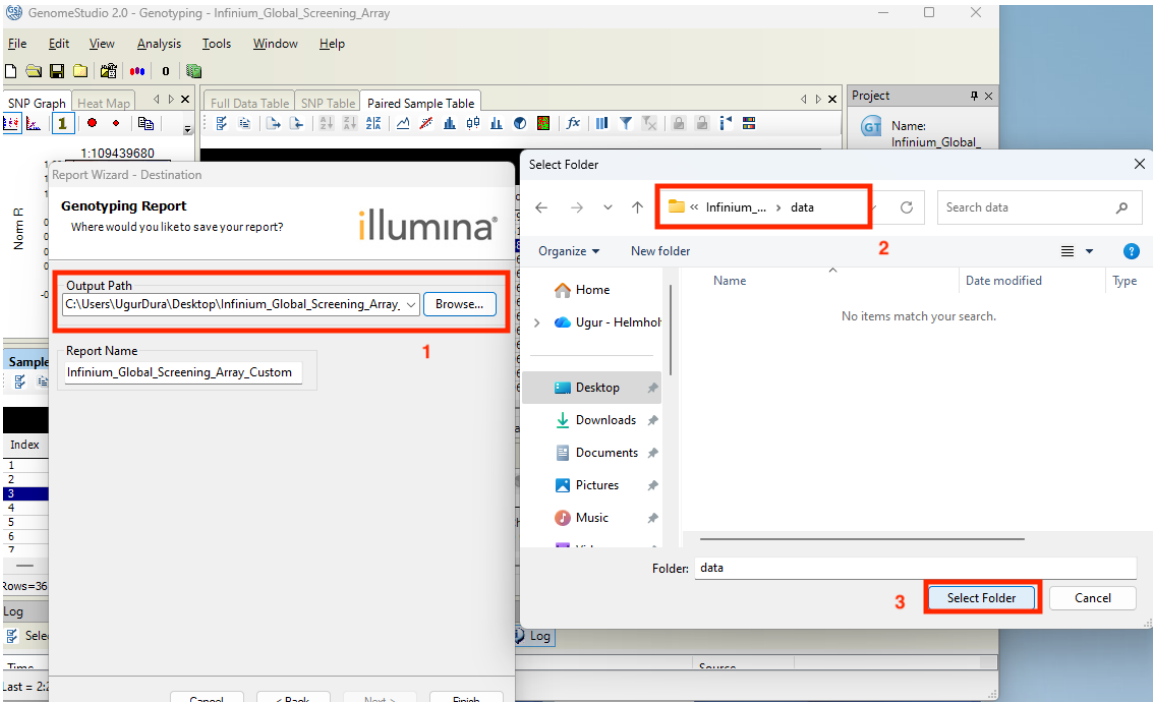

Figure 3.29: Configuring output path and report name. Choose the `data/` directory as the output path to keep all exported files organized in one location.

3.5.7 PLINK Export Progress

1. After clicking Finish, the PLINK export process will begin (see Figure 3.30)
2. A progress bar will appear showing the export status
3. Wait for the process to complete
4. This typically takes 2-5 minutes for 36 samples

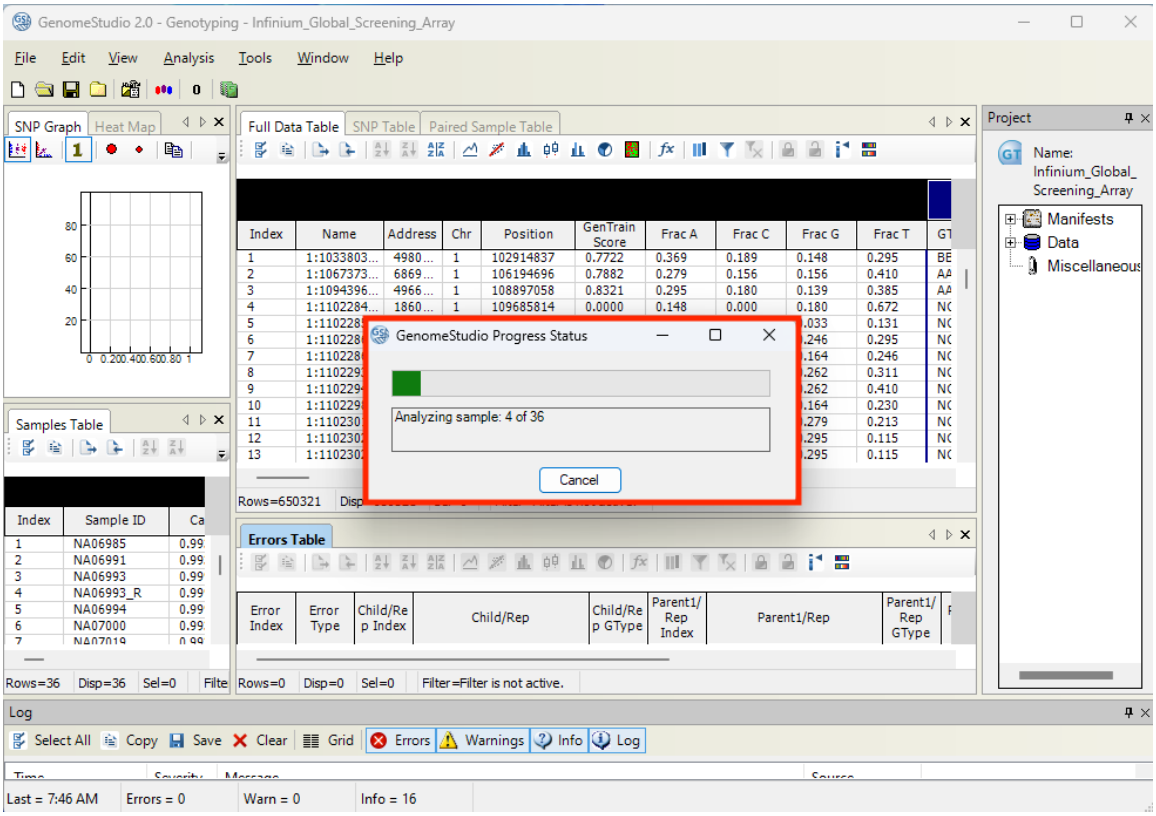

Figure 3.30: PLINK file generation in progress. The progress bar shows the current status of the export process. Wait for it to complete before proceeding.

3.5.8 Verify PLINK Output Files

After the PLINK export completes successfully, navigate to the `data/` directory to verify the output files (see Figure 3.31).

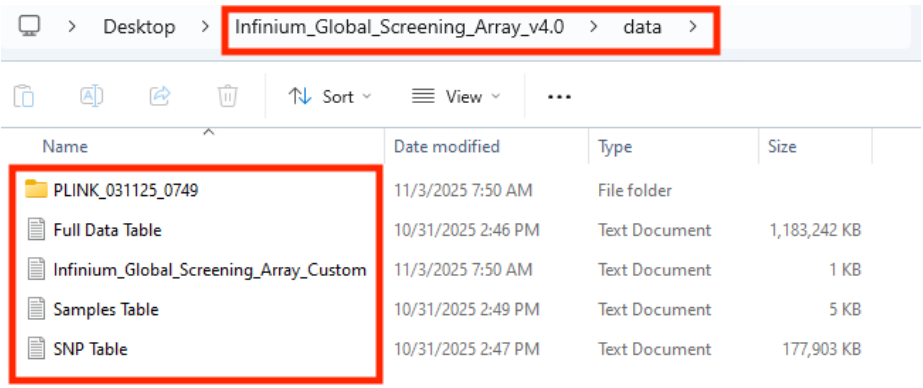

Figure 3.31: Contents of the `data/` directory after PLINK export completion. The directory contains the PLINK output folder, exported tables, and the custom report file.

You should see the following items in the `data/` directory:

File/Folder	Type	Description
`PLINK_031125_0749/`	Folder	Contains all PLINK output files (.ped, .map, .bed, .bim, .fam)
`Full_Data_Table.txt`	Text File	Full Data Table export (∼1.2 GB)
`SNP_Table.txt`	Text File	SNP Table export
`Samples_Table.txt`	Text File	Samples Table export (∼5 KB)
`Infinium_Global_Screening_Array_Custom.txt`	Text File	PLINK report file (∼1 KB, not needed for pipeline)

Table 3.2: Expected contents of the `data/` directory after all exports

PLINK Output Folder

Important Notes:

- The PLINK output folder name includes a timestamp (e.g., `PLINK_031125_0749`)
- Inside this folder, you will find the PLINK format files (.ped, .map, .bed, .bim, .fam)
- The Digital Karyotyping Pipeline will use files from this PLINK folder
- The `Infinium_Global_Screening_Array_Custom.txt` report file is not required for the pipeline

4. Final Data Organization

After completing all exports from GenomeStudio software (Illumina), you need to organize your files and prepare additional required files for the Digital Karyotyping Pipeline.

4.0.1 Organize Exported Data

Your `data/` directory should now contain all the exported files from Genome Studio (see Figure 4.1):

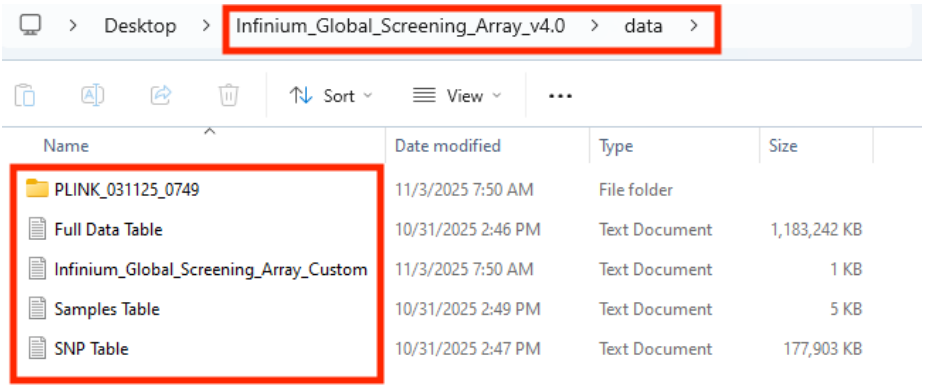

Figure 4.1: Contents of the `data/` directory after completing all Genome Studio exports including PLINK files. This directory contains the Full Data Table, SNP Table, Samples Table, PLINK output folder (with timestamp), and the custom report file.

4.0.2 Prepare Additional Required Files

In addition to the Genome Studio exports, the Digital Karyotyping Pipeline requires several additional files:

Manifest File (CSV Format)

The pipeline requires the manifest file in CSV format (not BPM format):

- **File**: `GSA-48v4-0_20085471_D2.csv`
- **Location**: Already downloaded in Chapter 2
- **Action**: Copy this file to your `data/` directory
- **Source**: `https://support.illumina.com/array/array_kits/infinium-global-screening-array/downloads.html`

Reference Genome (FASTA)

The pipeline requires the human reference genome in FASTA format:

- **File**: `Homo_sapiens.GRCh38.dna.primary_assembly.fa`
- **Build**: GRCh38 (must match the manifest file build)
- **Source**: Available from Ensembl or NCBI
- **URL**: `http://ftp.ensembl.org/pub/release-110/fasta/homo_sapiens/dna/`
- **Action**: Download and place in a reference genome directory

Pseudoautosomal Region (PAR) Coordinates

The pipeline requires PAR coordinates for correct sex chromosome analysis:

- **File**: `PAR_Coord_GRCh38.txt`
- **Location**: Included in the pipeline repository
- **Path**: `pipeline_digital_karyotyping/datasets/PAR/PAR_Coord_GRCh38.txt`
- **Action**: No download needed, already in the repository

Sample Sheet File

The pipeline requires a sample sheet file to specify which samples to compare:

- **Format**: Tab-separated or Excel file with two columns: `Sample` and `Reference`
- **Purpose**: Defines sample-to-reference comparisons for CNV detection
- **Example**: Self-comparisons (sample vs. itself) for baseline analysis

Create a file named `sample_sheet.txt` with the following format:

```
Sample          Reference
NA06985         NA06985
NA06991         NA06991
NA06993         NA06993
NA06993_R       NA06993_R
NA06993_R       NA06993
NA06994         NA06994
NA07000         NA07000
NA07019         NA07019
NA07029         NA07029
NA07056         NA07056
NA10838         NA10838
NA10846         NA10846
NA11832         NA11832
NA11881         NA11881
NA11992         NA11992
NA11993         NA11993
NA11993_R       NA11993_R
NA11993_R       NA11993
NA11995         NA11995
NA12003         NA12003
NA12003_R       NA12003_R
NA12003_R       NA12003
NA12044         NA12044
NA12155         NA12155
NA12156         NA12156
NA12156_R       NA12156_R
NA12156_R       NA12156
NA12239         NA12239
NA12248         NA12248
NA12248_R       NA12248_R
NA12248_R       NA12248
NA12249         NA12249
NA12760         NA12760
NA12812         NA12812
NA12813         NA12813
NA12875         NA12875
NA12877         NA12877
NA12877_R       NA12877_R
NA12877_R       NA12877
NA12878         NA12878
NA12891         NA12891
NA12892         NA12892
```

Sample Sheet Explanation

Understanding Sample Pairing:

- **Self-comparisons** (e.g., NA06985 vs NA06985): Baseline analysis to identify inherent CNVs
- **Replicate comparisons** (e.g., NA06993_R vs NA06993): Quality control to verify reproducibility
- **Sample IDs must match** those in the Genome Studio data exactly
- The file can be in `.txt`, `.tsv`, or `.xls` format

4.0.3 Final Directory Structure

Organize all files in a comprehensive directory structure:

```
Infinium_Global_Screening_Array_v4.0/
|-- idats_Demo_36/
|   `-- ... (original .idat files)
|-- GSA-48v4-0_20085471_D2.bpm
|-- GSA-48v4-0_20085471_D2_ClusterFile.egt
|-- GSA-48v4-0_D2_SampleSheet_Demo_36.csv
|-- GenomeStudio_Project/
|   `-- GSA_Demo_36_Samples.bsc (Genome Studio project file)
`-- data/
    |-- Full_Data_Table.txt (~1.2 GB)
    |-- Samples_Table.txt (~5 KB)
    |-- SNP_Table.txt
    |-- PLINK_031125_0749/ (PLINK output folder with timestamp)
    |   |-- ... (.ped, .map, .bed, .bim, .fam files)
    |-- Infinium_Global_Screening_Array_Custom.txt (report file)
    |-- GSA-48v4-0_20085471_D2.csv (Manifest CSV - REQUIRED)
    `-- sample_sheet.txt (or .xls) (Sample sheet - REQUIRED)

Reference_Files/ (separate directory, can be anywhere)
|-- Homo_sapiens.GRCh38.dna.primary_assembly.fa
`-- PAR_Coord_GRCh38.txt (from pipeline repository)
```

4.0.4 Complete File Checklist

Before configuring the pipeline, verify you have all required files:

Category	File Type	File Name/Description	Status
Genome Studio	Full Data Table	Full_Data_Table.txt	☐
	Samples Table	Samples_Table.txt	☐
	SNP Table	SNP_Table.txt	☐
PLINK Files	PLINK Folder	PLINK_[timestamp]/ directory	☐
	PED File	*.ped (inside PLINK folder)	☐
	MAP File	*.map (inside PLINK folder)	☐
	BED File	*.bed (inside PLINK folder)	☐
	BIM File	*.bim (inside PLINK folder)	☐
	FAM File	*.fam (inside PLINK folder)	☐
Additional	Manifest (CSV)	GSA-48v4-0_20085471_D2.csv	☐
	Reference Genome	Homo_sapiens.GRCh38...fa	☐
	PAR Coordinates	PAR_Coord_GRCh38.txt	☐
	Sample Sheet	sample_sheet.txt/.xls	☐

Table 4.1: Complete checklist of all required files for the Digital Karyotyping Pipeline

5. Configuring the Pipeline

Now that all data has been exported and organized, you can configure the Digital Karyotyping Pipeline to analyze the demo dataset.

5.0.1 Understanding the params.yaml File

The Digital Karyotyping Pipeline uses a YAML configuration file (`params.yaml`) to specify input files and analysis parameters. The template is located at:

`pipeline_digital_karyotyping/templates/params.yaml`

5.0.2 Configuration Steps

1. **Copy the template**:

```
cp templates/params.yaml params_demo_dataset.yaml

```

2. **Edit the configuration file** with your file paths:

```
nano params_demo_dataset.yaml

```

5.0.3 Parameter Configuration

Update the following parameters in `params_demo_dataset.yaml` with the correct file paths:

Reference Genome Files

```
# Reference Genome (GRCh38 for this demo dataset)
PAR: /path/to/pipeline_digital_karyotyping/datasets/PAR/PAR_Coord_GRCh38.txt

fasta: /path/to/Reference_Files/Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

Important: Genome Build

Ensure you use **GRCh38** reference files to match the manifest file (GSA-48v4-0 Build 38). Using mismatched genome builds will cause errors in the analysis.

Project Information

```
# Project Information
project_ID: 'GSA_Demo_36_Samples'
responsible_person: 'Your_Name'
```

Sample Sheet File

```
# Sample Sheet File
samples_refs: /path/to/Infinium_Global_Screening_Array_v4.0/data/sample_sheet.txt
```

GenomeStudio software (Illumina) Export Files

```
# Illumina Manifest (CSV format)
manifest: /path/to/Infinium_Global_Screening_Array_v4.0/data/GSA-48v4-0_20085471_D2.csv

# GenomeStudio Export Files
fullTable: /path/to/Infinium_Global_Screening_Array_v4.0/data/Full_Data_Table.txt

samplesTable: /path/to/Infinium_Global_Screening_Array_v4.0/data/Samples_Table.txt

snpTable: /path/to/Infinium_Global_Screening_Array_v4.0/data/SNP_Table.txt
```

PLINK Files Directory

```
# PLINK Files (directory containing .ped, .map, .bed, .bim, .fam)
gsplink: /path/to/Infinium_Global_Screening_Array_v4.0/data/PLINK_031125_0749
```

Critical: PLINK Directory Path

Important! The `gsplink` parameter should point to the PLINK output **directory** (the folder with the timestamp, e.g., `PLINK_031125_0749`), NOT to individual .ped/.map files. The pipeline will automatically find all PLINK files inside this directory.

Output Directory (Optional)

```
# Output Directory (optional - can also use --outdir flag)
outdir: '/path/to/results/GSA_Demo_36_Results'
```

5.0.4 Complete params.yaml Example for Demo Dataset

Here is a complete example configuration for the Illumina demo dataset:

```
# Reference Genome (GRCh38)
PAR: /home/user/pipeline_digital_karyotyping/datasets/PAR/PAR_Coord_GRCh38.txt
fasta: /home/user/Reference_Files/Homo_sapiens.GRCh38.dna.primary_assembly.fa

# Project Information
project_ID: 'GSA_Demo_36_Samples'
responsible_person: 'Your_Name'

# Sample Sheet File
samples_refs: /home/user/Infinium_Global_Screening_Array_v4.0/data/sample_sheet.txt

# Illumina Manifest (CSV format)
manifest: /home/user/Infinium_Global_Screening_Array_v4.0/data/GSA-48v4-0_20085471_D2.csv

# GenomeStudio Export Files
fullTable: /home/user/Infinium_Global_Screening_Array_v4.0/data/Full_Data_Table.txt
samplesTable: /home/user/Infinium_Global_Screening_Array_v4.0/data/Samples_Table.txt
snpTable: /home/user/Infinium_Global_Screening_Array_v4.0/data/SNP_Table.txt

# PLINK Files Directory
gsplink: /home/user/Infinium_Global_Screening_Array_v4.0/data/PLINK_031125_0749

# Output Directory
outdir: '/home/user/results/GSA_Demo_36_Results'
```

Ready to Run the Pipeline

Configuration Complete! Once you have verified all file paths and parameters, you are ready to run the Digital Karyotyping Pipeline with the demo dataset. The pipeline will:

- Process CNV data from LRR and BAF values
- Generate quality control metrics
- Detect copy number variations
- Create an interactive KaryoPlayground report

Refer to the main pipeline documentation for instructions on running the analysis.

5.0.5 Next Steps: Running the Pipeline

With all data extracted and configured, you are now ready to run the Digital Karyotyping Pipeline:

For more information on running and interpreting the Digital Karyotyping Pipeline results, refer to the main pipeline documentation.

6. Expected Pipeline Results

After successfully running the Digital Karyotyping Pipeline with the Illumina demo dataset, you will generate an interactive KaryoPlayground report with comprehensive visualizations and analysis results.

The pipeline produces an HTML-based interactive report that includes the KaryoPlayground home page, LRR/BAF plots, and genome-wide karyograms. Figure 6.1 shows examples of the expected output when running the pipeline with the demo dataset.

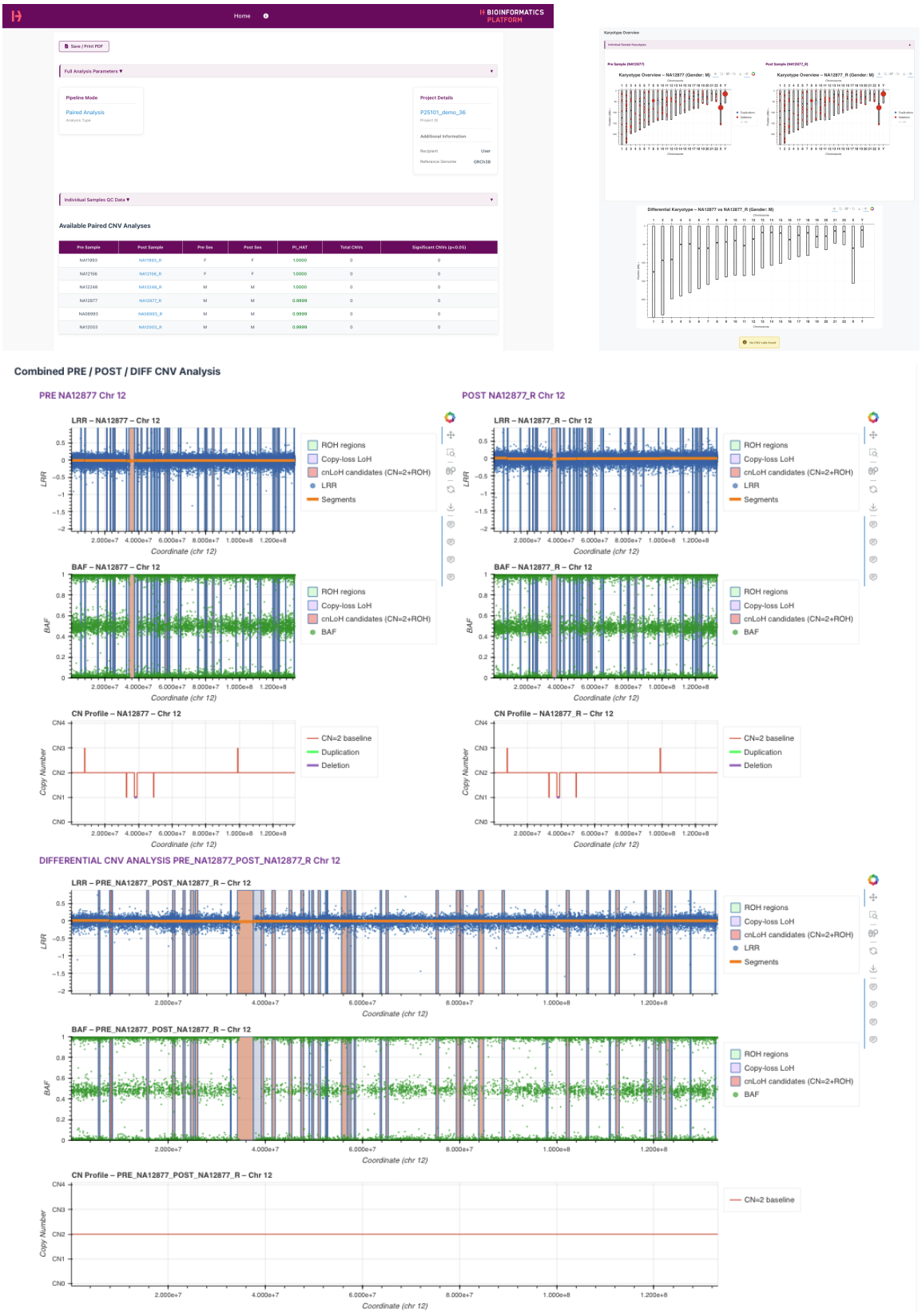

Figure 6.1: Expected pipeline results when running the Digital Karyotyping Pipeline with the Illumina demo dataset. Top left: KaryoPlayground home page with interactive dashboard. Top right: Genome-wide karyograms displaying CNV patterns across all chromosomes with color-coded copy number gains (red/orange) and losses (blue/green). Bottom: LRR and BAF plots showing pre-processing and post-processing results for CNV detection. Source: Illumina 2024 Infinium Global Screening Array v4.0 Used under license from Illumina, Inc. All Rights Reserved.

References

This guide references the following Illumina documentation, software tools, and scientific publications related to microarray genotyping, data analysis, and digital karyotyping:

[1] Illumina, Inc., *Genomestudio genotyping module v2.0 software guide*, Document # 11319113 v01, Illumina, Inc., San Diego, CA, USA, 2020 (cited on page 5).

[2] Illumina, Inc., *Plink input report plug-in v2.1.4 for genomestudio*, Illumina, Inc., San Diego, CA, USA, 2019 (cited on pages 6, 11).

[3] S. Purcell et al., "Plink: A tool set for whole-genome association and population-based linkage analyses", *The American Journal of Human Genetics*, volume 81, number 3, pages 559–575, 2007. DOI: 10.1086/519795. [Online]. Available: https://doi.org/10.1086/519795 (cited on pages 6, 11).

[4] C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation plink: Rising to the challenge of larger and richer datasets", *GigaScience*, volume 4, number 1, page 7, 2015. DOI: 10.1186/s13742-015-0047-8. [Online]. Available: https://doi.org/10.1186/s13742-015-0047-8 (cited on pages 6, 11).

[5] Illumina, Inc., *Infinium global screening array-24 v4.0 beadchip data sheet*, Illumina, Inc., San Diego, CA, USA, 2024. [Online]. Available: https://www.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html (cited on page 7).

[6] Illumina, Inc., *Infinium global screening array v4.0 support files*, Accessed: 2024, Illumina, Inc., 2024 (cited on page 7).

[7] K. L. Gunderson, F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee, "A genome-wide scalable snp genotyping assay using microarray technology", *Nature Genetics*, volume 37, number 5, pages 549–554, 2005. DOI: 10.1038/ng1547. [Online]. Available: https://doi.org/10.1038/ng1547 (cited on page 7).

[8] F. J. Steemers, W. Chang, G. Lee, D. L. Barker, R. Shen, and K. L. Gunderson, "Whole-genome genotyping with the single-base extension assay", *Nature Methods*, volume 3, number 1, pages 31–33, 2006. DOI: 10.1038/nmeth842. [Online]. Available: https://doi.org/10.1038/nmeth842 (cited on page 7).

[9] Genome Reference Consortium, "Human build 38 patch release 14 (grch38.p14)", *Genome Reference Consortium*, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40 (cited on page 7).

[10] D. A. Peiffer et al., “High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping”, *Genome Research*, volume 16, number 9, pages 1136–1148, 2006. DOI: 10.1101/gr.5402306. [Online]. Available: https://doi.org/10.1101/gr.5402306 (cited on pages 11, 26).

[11] S. J. Diskin et al., “Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms”, *Nucleic Acids Research*, volume 36, number 19, e126, 2008. DOI: 10.1093/nar/gkn556. [Online]. Available: https://doi.org/10.1093/nar/gkn556 (cited on page 11).

[12] R. McQuillan et al., “Runs of homozygosity in european populations”, *The American Journal of Human Genetics*, volume 83, number 3, pages 359–372, 2008. DOI: 10.1016/j.ajhg.2008.08.007. [Online]. Available: https://doi.org/10.1016/j.ajhg.2008.08.007 (cited on page 11).