Auto-qPCR: A Python based web app for automated and reproducible analysis of qPCR data

Gilles Maussion[*1,2], Rhalena A. Thomas*[*1], Iveta Demirova[1], Gracia Gu[1], Eddie Cai[1], Carol X.-Q Chen[1], Narges Abdian[1], Theodore J.P. Strauss[3], Sabah Kelai[2], Angela Nauleau-Javaudin[1], Lenore K. Beitel[1], Nicolas Ramoz[2], Philip Gorwood[2], Thomas M. Durcan[1]*

[1] The Neuro's Early Drug Discovery Unit (EDDU), McGill University, 3801 University Street, Montreal, QC Canada H3A 2B4

[2] INSERM U1266, Institute of Psychiatry and Neuroscience of Paris, Paris, France

[3] McConnell MNI Brain Imaging Center, McGill University, 3801 University Street, Montreal, QC Canada H3A 2B4

* These authors have contributed equally

§ Corresponding author:

Thomas M. Durcan, Ph.D.

McGill University

The Neuro's Early Drug Discovery Unit (EDDU),

3801 University St., Montreal, QC H3A 2B4 Canada

Mail: thomas.durcan@mcgill.ca

Phone: 514-398-6933

Running Title: Auto-qPCR, a web app for qPCR analysis

Keywords: qPCR, absolute quantification, relative quantification, bioinformatics, data processing, normalization, differential analyses, statistics, graphic representation, web app, genomic stability, differentiation, RNA, iPSC

Abbreviations: CT (cycle threshold), qPCR (quantitative polymerase chain reaction), iPSC (induced pluripotent stem cells), CNVs (copy number variants), SNVs (single nucleotide variants), DA (dopaminergic), Neural Precursor Cells (NPC), DA neurons (DANs)

Figures – 5

Supplementary Tables-16

Abstract

Quantifying changes in DNA and RNA levels is an essential component of any molecular biology toolkit. Quantitative real time PCR (qPCR) techniques, in both clinical and basic research labs, have evolved to become both routine and standardized. However, the analysis of qPCR data includes many steps that are time consuming and cumbersome, which can lead to mistakes and misinterpretation of data. To address this bottleneck, we have developed an open source software, written in Python, to automate the processing of csv output files from any qPCR machine, using standard calculations that are usually performed manually. Auto-qPCR is a tool that saves time when computing this type of data, helping to ensure standardization of qPCR experiment analyses. Unlike other software packages that process qPCR data, our web-based app (http://auto-q-pcr.com/) is easy to use and does not require programming knowledge or software installation. Additionally, we provide examples of four different data processing modes within one program: (1) cDNA quantification to identify genomic deletion or duplication events, (2) assessment of gene expression levels using an absolute model, (3) relative quantification, and (4) relative quantification with a reference sample. Auto-qPCR also includes options for statistical analysis of the data. Using this software, we performed analysis of differential gene expression following an initial data processing and provide graphs of the findings prepared through the Auto-qPCR program. Thus, our open access Auto-qPCR software saves the time of manual data analysis and provides a more systematic workflow, minimizing the risk of errors when done manually. Our program constitutes a new tool that can be incorporated into bioinformatic and molecular biology pipelines in clinical and research labs.

Introduction

Polymerase chain reaction (PCR) is a temperature cycle-based DNA polymerization technique that helps identify a nucleic acid fragment of interest by increasing its proportion relative to others (Saiki et al., 1985). Initially the technique was primarily used to visualize DNA fragments for cloning (Scharf et al., 1986; Magnuson et al., 1996) or genotyping (Saiki et al., 1986; Mullis and Faloona, 1987; Beggs et al., 1990), but can now be used to investigate genetic polymorphisms and mutations (Ye et al., 2001; De la Vega et al., 2005), copy number variants (CNVs) (D'Haene et al., 2010), single nucleotide variants (SNVs), point mutations, and genetic deletion/duplication events (Charbonnier et al., 2000). With the development of fluorogenic probes and dyes capable of binding newly synthesized DNA, PCR became more quantitative, leading to innovative tools for quantifying relative transcript levels for one or more genes, now referred to as quantitative PCR or qPCR. With these technological advancements, qPCR is now used to quantify messenger RNA (mRNA) (Wong and Medrano, 2005), long non-coding RNA (Gupta et al., 2010), and microRNAs (Shi and Chiang, 2005; Varkonyi-Gasic et al., 2007). Preceded by chromatin immunoprecipitation, DNA-protein interactions (Mukhopadhyay et al., 2008) or epigenetic modifications (Dahl and Collas, 2007; Milne et al., 2009) at specific DNA loci can be detected through qPCR. qPCR has also become an essential tool, routinely used in clinics for diagnosis, and in research labs. For instance, with induced pluripotent stem cell (iPSC) research, qPCR provides a critical test to assess the genomic stability of the iPSCs at key hotspot sites (Artyukhov et al., 2017; Yoshihara et al., 2017) . Thus, the advent of PCR has revolutionized our ability to analyze and quantify nucleic acids and has made qPCR a standard technique.

Although qPCR experiments are already automated at the data acquisition stage, with thermocycler software providing “by default” pre-processing procedures (Pabinger et al., 2014) , data analysis is still highly time consuming and error prone, especially when processing large numbers of data points. The user must intervene at multiple steps: raw data exclusion, identification of outliers, normalization(s) and differential expression analyses. Without guidelines or standardized procedures, such manual analysis can potentially introduce “user-dependent” variation and errors. To both simplify and accelerate this data analysis step for qPCR datasets, we have created a Python-based, open source, user-friendly web application “Auto-qPCR” to process exported qPCR data and to provide visual representations of the data, with accompanying statistical analysis. The program can be found at the website http://auto-q-pcr.com/ and can treat csv files from exported qPCR datasets in line with the two commonly used molecular biology approaches: (i) absolute quantification where all estimations rely on orthogonal projection of the samples of interest onto a calibration curve (Bustin, 2000), and (ii) relative quantification that relies on difference of cycle threshold (CT) values between the gene of interest and endogenous controls (Pfaffl, 2001).

In this manuscript we present datasets generated from qPCR experiments that illustrate four distinct computational modes to assess the presence of deletion or duplication events in DNA (genomic instability) and to quantify normalized RNA expression levels across several experiments. In addition, statistical analyses and graphs of the findings are generated by Auto-qPCR. Together, Auto-qPCR provides an all-in-one solution for the user, going from datasets to graphs, all within one web-based software package. While other open source qPCR analysis software programs and web apps (Rancurel et al., 2019; Zanardi et al., 2019; Krahenbuhl et al., 2020) are available, they are only able to normalize, compare and display qPCR data generated with one of the two models of quantification (Bustin, 2000; Pfaffl, 2001). In contrast, Auto-qPCR provides a comprehensive data analysis package for qPCR experiments. Using the web app does not require any programming knowledge, account creation or desktop installation. Additionally, the program has been designed to assist the user at each step of the analysis, once the exported data files have been collected from the qPCR system.

Auto-qPCR can be used to analyse qPCR data in a reproducible manner, simplifying data analysis, avoiding potential human error and saving time. In this manuscript, we describe some of the uses of the software and outline the steps required from entering an individual dataset to complete statistical analysis and graphical presentation of the data.

Material and Methods

Culture of iPSC lines

In order to illustrate the four different models of quantification by qPCR, managed by the Auto-qPCR program, we used 11 different iPSC cells lines described in **Table S1**. Briefly, the GM25953, GM25974, GM25975 and GM25952 IPS cell lines were obtained by reprogramming of fibroblasts using episomal vectors. NCRM1 is an iPSC line generated by episomal reprogramming of CD34+ cord blood cells and was obtained from the NIH. The AJG001-C4 iPSC were reprogrammed from PBMCs using episomal vectors. The other iPSC cells lines were reprogrammed with transducing retrovirus from PBMCs (AIW001-2; AIW002-2), fibroblasts (AJC001-5, KYOU-DRX0190B) or lymphocytes (522-2666-2). Quality control profiling for the iPSCs used was outlined in a previous study (Chen et al., 2021).

The iPSCs were seeded on Matrigel-coated dishes and expanded in mTESR1 (Stemcell Technologies) or Essential 8 (ThermoFisher Scientific) media. On the first day of culture, cells were seeded at 10 to 15% confluency and incubated at 37°C in a 5% CO_2 environment. The media was changed daily until the cultures reached 70% confluency. Cells harbouring irregular borders, or transparent centres were manually removed from the dish prior to dissociation with Gentle Cell Dissociation media (Stemcell Technologies) for 6 minutes at room temperature to obtain small aggregates of colonies. The IPSCs were then seeded and differentiated into cortical or dopaminergic neuronal progenitors or neurons.

Generation of cortical and dopaminergic neurons

The induction of cortical progenitors was performed as described previously (Bell et al., 2017). The neural progenitor cells were dissociated, then purified by culturing in suspension for 48 hours in expansion medium: (DMEM/F12 with Glutamax supplemented with N2, B27 (ThermoFisher Scientific), NEAA (Gibco), laminin (1µg/ml) (Sigma) plus the growth factors EGF (20 ng/ml) and FGF (20 ng/ml) (Peprotech). The media used for cortical differentiation is described in the standard operating procedure published on the Early Drug Discovery Unit (EDDU) website (Chen et al., 2021) Briefly, purified NPCs were replated on Matrigel-coated dishes (Thermo-Fisher). Once cells attained 100% confluency, NPCs were passaged and seeded on a Poly-Ornithine-laminin coated dishes to be differentiated into neurons. Cells were switched for 24 hours to 50% Neurobasal (NB) medium, and 24 hours later placed in 100% NB medium with AraC (0.1µM) (Sigma) to reduce levels of dividing cells. After the third day of differentiation, cells were maintained in 100% NB medium without AraC for four days before being collected in lysis buffer for RNA extraction. IPSCs were induced into dopaminergic NPCs (DA-NPCs) according to methods previously described (Kriks et al., 2011), modified according to methods used within the group (Chen et al., 2019b). DA-NPCs were subsequently differentiated into dopaminergic

neurons (DANs), with immunostaining and qPCR analysis performed at four and six weeks of maturation from the NPC stage (Chen et al., 2019a).

Cell collection for DNA or RNA extraction

IPSCs were dissociated with Gentle Cell Dissociation Reagent (Stem Cell Technologies) while Accutase® Cell Dissociation Reagent (Thermo Fisher Scientific) was used to dissociate NPCs and iPSC-derived neurons. After 5 minutes incubation at 37°C with the indicated dissociation agent, cells were collected and harvested by centrifugation for 3 minutes at 1200 rpm. Cell pellets were resuspended in lysis buffer and stored at -80°C before DNA or total RNA extraction with the Genomic DNA Mini (Blood/Culture Cell) (Genesis) or mRNAeasy (Qiagen) kits, respectively.

cDNA synthesis and Quantitative PCR

Reverse transcription reactions were performed on 400ng of total RNA extract to obtain cDNA in a 40 µl total volume containing, 0.5µg random primers, 0.5mM dNTPs, 0.01M DTT and 400 U/µl-MMLV RT (Carlsbad, CA, USA). The reactions were conducted in singleplex, in a 10µl total volume containing 2X Taqman Fast Advanced Master Mix (5µl), 20X Taqman primers/probe set (0.5µl) (Thermo Fisher Scientific), 1µl of diluted cDNA and H_2O. Real-time PRC (RT-PCR) was performed on a QuantStudio 3 machine (Thermo Fisher Scientific). Primers/probe sets used were from Applied Biosystems and selected from the assays available on Thermo Fisher Scientific web site. The primers/probe sets were chosen to cover the most important number of alternative transcripts for a given gene. Two endogenous controls (beta-actin and GAPDH) were used for normalization (**Table S2)**. With the exception of the assay for *GAPDH*, the amplicons overlapped two exons, avoiding amplification of genomic DNA that could remain from incomplete DNAse digestion. A refseq sequence used for designing the primer/probe set assay has also been reported.

Collection of external data set

An external qPCR data set was provided from an earlier published study (Kelai et al., 2008), which quantified levels of *Nrxns* and *Nlgn* transcripts in the subcortical areas of the brains from mice submitted to conditioned place preference (CPP) with cocaine. Briefly, the mouse brain was sectioned with a cryostat (Leica CM3050S) and subcortical areas (subthalamic nucleus, globus pallidum and substantia nigra) isolated by laser capture microdissection (Leica ASLMD instrument with LMD 5.0 and IM1000 software (Leica). RNA was extracted with the Arcturus PicoPure kit and reverse transcription performed as described above. The qPCR experiments

were performed according to an absolute quantification design on the Opticon 2 PCR machine (Biorad) connected to the Opticon monitor 2 software. B2Microglobulin (*B2M*) was used as endogenous control. For the current manuscript, data were re-extracted from the Opticon monitor 2 files as csv files and analyzed by Auto-qPCR.

Program development

The program was written in Python using Pandas and NumPy. The structure consists of a script main.py which reads in and formats the raw csv data and calls the next scripts. The script Auto-qPCR.py processes the data (transforms data, identifies controls and removes outliers) and then calls the model selected in the web-interface to quantify the RNA concentration. The models are contained within separate scripts (absolute.py, relative.py and stability.py). The relative model with two normalizations (delta-delta CT) is the same calculation used for the genomic instability test. The program calls the script 'stability.py' when either the stability or relative dd-CT is selected by the user. The processed results with all samples (separate technical replicates) and a summary (mean values) spreadsheet is created and saved in csv format.

We also provide options within the program to perform statistics for Student's t-test, one-way and two-way ANOVAs followed by multiple-t-test with FDR correction and equivalent non-parametric test. The scripts were written in Python using Pingouin, Scipy and Pandas and can be found in statistics.py. The user inputs the number of groups for multiple comparisons based on the annotation present in the input data. The program will format the data, run the selected statistical model and then provide a results spreadsheet in the form of a csv file for the t-tests, a csv file for the ANOVA results and a separate csv file for the posthoc test results. All the output tables and plots are combined into a zip file that the end user can download from the website. All the input and output data are cleared after processing and no user data is stored in the web app. **Table S3** summarizes the organization and function of the script files for the program.

Program function - input data processing and quantification

The Auto-qPCR program reads the raw data in the form of csv files (comma-delimited) from the PCR machine and reformats it into a data frame in Python. The csv files are selected from the dropdown arrow that allows the user to find files using their computer's file navigator. The user enters information into the web-app to match the experimental design of the data to be analyzed and these are read as arguments by the software. See **Table S4** for a list of all the user inputs. The raw data in the form of a csv file is read into Python by P searching for the first line with multiple column names and a data frame is created containing all the raw data and sample information. The values for the reference genes/targets (*GAPDH*, *ACTB*) are calculated for each

sample and technical replicate (cell line, time point, treatment condition) separately. To detect outliers, the standard deviation (std) of the technical replicates for a given sample is calculated, if the std is greater than the cut-off (the default value is 0.3), then the technical replicate furthest from the sample mean is removed. The process occurs recursively until the std is less than the cut-off or only the value of "max outliers" is reached. Max outliers is set to 0.5 by default which means that outliers will be removed until two technical replicates remain. The number of technical replicates included in each final sample mean is indicated in the 'summary_data.csv' output file. The 'preserve highly variable replicates' option adds a second criteria before a replicate is removed. If the CT-std is less than 0.3, but the absolute (mean-median)/median is less than 0.1, replicates are preserved. This helps to account for a lack of a clear outlier, where two of three replicates are close to equally distributed around the mean. The next processing steps are model dependent: **The absolute model** calculates the ratio between the gene of interest and each control. For each gene/target of interest the normalized value is calculated against the mean of each control target separately, then the mean value from normalized to controls is calculated. **The relative model ΔCT (delta CT),** without a calibration sample, calculates the ΔCT by subtracting the log2 control CT value from the log2 CT value for the target for each control and then takes the mean value of the resulting deltas. **The relative model ΔΔCT (delta-delta CT)**, with a calibration sample and the **genomic stability model**, individually calculates the ΔCT for the target in test sample and the reference/calibration sample then calculates the ΔΔCT by subtracting the reference ΔCT from the test sample. For all models the mean value of technical replicates is calculated for each target.

For the relative models the values for the reference genes are calculated separately for each csv file. The data from one csv file will not be applied to another csv file. For the absolute model, qPCR output for each gene is found in a separate csv file and the selected endogenous controls will be applied to all the data input in one analysis. For all models, two spreadsheets are outputted as csv (comma-delimited) files that can be opened in Excel, LibreOffice or any text editor. The user will receive "clean_data.csv" where delta CT is calculated for every technical replicate, the outliers are included and indicated by "TRUE" in the column "Outlier". The summary output contains the mean, standard deviation (std) and standard error (SE)for each sample technical replicates. The table of the data, "summary_data.csv", is appropriate for further analysis in another statistical program (R, SASS, Prism).

Program function – statistical analysis

For testing differential gene expression, the user selects the statistic option and files in a form to indicate the conditions of the experiment. Either paired test (t-test) or multiple comparisons (one-way ANOVA or 2-way ANOVA) to investigate interaction effects is selected. The number of groups to compare is inputted by the user, if a two-way ANOVA is selected the total number of conditions is entered. The names of the variables to

be grouped by must be within either the 'sample names' column of the raw csv data or an additional column (which is created when entering the experimental design into the thermal cycler). Users can also add a column to their data after exporting the csv file, although this will add a risk of copy/paste errors and add additional time to the analysis process. The user selects using independent or repeated measures as well as the normality of the distribution of the data, then the appropriate statistical test will be applied. See **Table S5** for the list of which analysis is applied for each setting. All default setting are maintained for statistical functions (for details see the Pingouin documentation at https://pingouin-stats.org/, the output has been reformatted to be more easily read and interpreted by users and for consistency across statistical outputs.

Program function – visualization

The plotting scripts were written using the Matplotlib bar chart function. The labels and axis settings were all adjusted directly within the script. If users want to change the visualization, they can do so in the plot.py script on a local server. The labels and colours cannot be adjusted within the web app. The user can dictate the gene/target order and the sample order (cell lines, treatments, time points) in the web app by entering the orders into the appropriate input box. The order variables (for example, cell lines or time points) appear in the grouped data for the summary plots and can also be set in the statistical analysis by the user. The user designates the variable order by entering the variable names in the web-form bar charts for each gene, for all genes grouped by sample, and for all samples grouped by genes (with and without endogenous controls) are automatically generated and saved as png files. If statistics are applied, two summary bar charts of the mean values are generated, grouped by the selected variable. For two-way ANOVA analysis, the summary bar chart will group the first variable on the x-axis and the second variable will be visualized in different colours and indicated in the legend.

Interface development and web app

The graphical user interface (GUI) was created using Flask, a package for integrating HTML and Python code. The GUI is written in JavaScript, CSS, HTML and Bootstrap4, a framework for building responsive websites. On our GitHub repository (https://github.com/neuroeddu/Auto-qPCR), we include the original command line version of the program and a version of the GUI that can be installed locally to run on a computer. A complete list of package dependencies and instructions to install and run the package app locally are posted in the GitHub repository. For users who wish to work offline, they may adapt any of the python scripts for their local server. The program was developed using git version control with multiple contributors. This study can be cited for analysis performed on the web app, local server installation and any incorporation and adaption of

our scripts in customized scripts. The web app is hosted by the Brain Imaging Centre at the Montreal Neurological Institute-Hospital (The Neuro) and was installed in a virtual machine directly from the public GitHub repository. When updates are available the changes will be applied to the web app using GitHub.

Data availability and reproducibility

All raw csv data files and output files used in plots are available at https://github.com/neuroeddu/Auto-qPCR, along with a user guide. The example input and output files in the paper are all available and organized by Figure names. The raw data used in the manuscript is found in the "Input Data" along with a document describing how the data was entered into the Auto-qPCR program "Notes_on_Datasets.docx". The example output folders are found under "Output Data" along with a description of the output files present in each folder. The "Notes_on_Datasets.docx" file contains all the parameters used in Auto-qPCR to obtain the plots in the manuscript from the "Input Data" used for each figure. The example output will be replicated identically if the same conditions are entered.

Illustrations

The schematic representation in **Figure 1** and simplified versions in **Figures 2-4** were created in Adobe Illustrator Creative Cloud 2020, with icons inserted from BioRender.

Results

The Auto-qPCR program functions with the workflow of a qPCR experiment

Auto-qPCR is a program conceived to process and analyse data generated using a qPCR thermo cycler. A qPCR experiment includes multiple steps that can be divided into two categories: (1) sample preparation to conduct the qPCR reaction, and (2) data analysis, visually represented in the schematic in **Figure 1**. For all experiments designed to quantify gene expression levels, RNA is extracted from biological samples and converted into cDNA. Similar extraction techniques are used to collect genomic DNA, whose amplification facilitates the detection of deletions, duplications or single nucleotide polymorphisms. DNA or cDNA are then combined with reagents for amplification of target regions proportional to the quantity of the original region of interest. Prior to performing qPCR *in vitro*, the user must generate the *in-silico* experimental layout using software that monitors the biochemical reaction. The user defines the experimental design (absolute or relative quantification), the method for detecting DNA synthesis (Taqman or SybrGreen) and the location of each sample within the plate. Finally, at the end of the qPCR process/cycle/program, the recorded data is exported and then would normally be analyzed manually.

We created Auto-qPCR to assist the user in data normalization, visualization and differential expression analyses. The program was designed for the most common uses of qPCR: detecting DNA fragment duplications or deletions, and quantifying gene expression levels according to the absolute or relative quantification models. With the aim of saving time for the user and avoiding copy-paste mistakes, missed numbers or inconsistency in application of data inclusion and inclusion rules, Auto-qPCR provides output datasheets of the processed data, graphical visualization of the data and statistical analysis based on the user defined experimental design. Here we provide examples of four different use cases for Auto-qPCR using real-life experimental datasets.

Genomic instability

A relatively new application for qPCR detects small changes within the genome from a deletion to a duplication of a DNA segment. DNA regions known to be highly susceptible to such events can be quantified using a genomic instability qPCR test. In the field of induced pluripotent stem cell (iPSC) research, genomic instability tests are critical for quality control to screen for these duplication/deletion events that can arise during reprogramming and prolonged cell passaging (Tosca et al., 2015; Yoshihara et al., 2017). We performed a qPCR test for genomic stability, where for each cell line, the signal from each DNA region of interest was compared to the endogenous control region. Next, each of the samples (cell lines GM25953, GM25975, GM25974, GM25952) were compared to a control sample of DNA, known to not have any changes within the regions that were tested for genomic instability.

We uploaded the data into the Auto-qPCR web app and selected the genomic instability model **(Figure 2A)**. For the endogenous control, we used a region on chromosome 4 (CHR4) as the target for normalization, a region of the genome known not to contain any instabilities, and for the reference sample we used DNA known not to have any instabilities as the calibrator, which we indicate as "normal" (**Figure 2A**). The genomic instability model has two steps of normalization for its general formula. This formula and the variables used in the example calculation are outlined in **Figure 2C**. First, the CT values from the control region (i.e. CHR4) for each cell line are subtracted from each region of interest. Next, the ΔCT from the "normal" DNA control is subtracted from the ΔCT calculated for each cell line sample. Finally, the mean is calculated from the average of multiple technical replicates included with the plate design for each sample. Thus, the ΔΔCT values are expressed as "Relative Quantification" according to the following formula: $RQ=2^{-\Delta\Delta CT}$. If the sample has no abnormalities (deletions or duplications) the values obtained should be equal or close to 1, except for targets in the X chromosome in a male individual in which the ratio would be expected to be at 0.5. As the DNA used for PCR amplification may come from a mixed population of cells, where only some cells carry a deletion or duplication, we set an acceptable range of variation as 0.3 above and below the expected value of 1: DNA regions with RQ values between that 0.7 and 1.3 are considered normal. Values below 0.7 indicate a deletion and values above 1.3 indicate an insertion. For ease of analysis, we have included a column in the output file from the Auto-qPCR program that indicates normal, insertion or deletion (**Table S6**). We found that all seven chromosomal regions in the four cell lines tested were between 0.7 and 1.3 and we concluded that no duplications or deletions were present (**Figure 2D and S1B**). Overall, we demonstrated how Auto-qPCR can be used to analyse the data from a genomic instability qPCR assay, and that the app effectively processed the data, creating a summary table and graph of the data.

Absolute Quantification

For absolute quantification experiments, the quantities of RNA transcripts for a gene of interest and the endogenous controls are first estimated with a calibration curve (**Figure 3A**) to provide a mathematical relationship between the CT values and the RNA concentration or quantity. The relationship is described by the equation $CT=a\log_2[RNA] + b$, where "a "is the slope and b is the Y-intercept (**Figure 3C**) (Ovstebo et al., 2003). The expression levels of the RNA molecule of interest are then given by the ratio of the estimated amount of RNA for a select transcript and the estimated amounts of endogenous controls (**Figure 3B**). Consequently, the values given as "Normalized Expression Levels" depend on the levels of transcript within the biological material used to set the calibration curves. The absolute quantification design is very powerful for comparing expression levels of a given gene of interest between two or more biological conditions or groups. We investigated the expression of 3 gene transcripts across six different cell lines (AIW001-2, AIW002-02, AJC001-5, 522-266-2, AJG001C4, NCRM1) at four different stages in the differentiation of neurons from

iPSCs. We wanted to determine if the cell lines acquired the desired phenotype based on the presence of specific dopaminergic and neuronal markers as they differentiated into DANs, and measured gene expression levels of *KCNJ6*, *SYP*, and *GRIA1,* markers of neuronal differentiation and synapse formation. We compared transcript levels in iPSCs with cells differentiated into DA-NPCDA, and NPC cultured in final DANs in neuronal differentiation media for 4 and 6 weeks (DA4W and DA6W). The calibration curve was made from a mix of the cDNAs generated from the reverse-transcribed RNA reactions from the four timepoints in the differentiation process and made of eight four-time serial dilutions to cover a linear relationship in a dynamic range from 1 to 16384-fold dilution (**Figure 3A**). Raw data was normalized with two endogenous controls (*GAPDH* and *βACTB*) (**Figure 3D to 3H**). We measured expression levels of the three transcripts (*KCNJ6*, *SYP* and *GRIA1*) in iPSCs, iPSC-derived DA-NPC and iPSC-derived DANs differentiated for four and six weeks, all across 6 control cell lines and calculated the normalized RNA levels using Auto-qPCR (**Figure 3D to 3H**). The Auto-qPCR app provides several graphical representations of the normalized expression data. The normalized expression data is shown as the mean of the technical replicates with error bars indicating the divergence observed between wells of a replicate for a given condition. The Auto-qPCR app will generate one bar chart for each gene measured; the output for *KCNJ6* is shown in **Figure 3D**. Two more bar charts were generated for each gene and sample observations plotted together (grouped by gene **Figure 3E** and by sample **Figure 3G**), allowing for an overview of the data and visualization of the biological variation between cell lines at a given stage.

To test for changes in gene expression over the different stages of neuronal differentiation, the cells lines were considered as six biological replicates for a given condition (**Figures 3E and 3G**). We have included a statistical module in our software that allows user-defined groups to be compared. We used the statistics option in the app to determine if there was significant differential expression of *KCNJ6*, *SYP* or *GRIA1* transcripts over the four developmental time points, treating cell lines as biological replicates. The groups to compare were defined as the differentiation time points (four groups) and considered as repeated measures. As there are more than two groups, the Auto-qPCR software runs a one-way-repeated measures ANOVA for each gene. Two summary plots (**Figure 3F and 3H**) and two statistical output tables were generated: one for the ANOVAs and one for the secondary measures (**Tables S7 and S8**). We found there was a significant effect of the differentiation stage on the expression of synaptic markers. The t-tests with false discovery rate (FDR) correction for pairwise comparisons of each stage showed that iPSCs have significantly less expression of each synaptic marker than DAN differentiated for 4 and 6 weeks (**Table S8**), indicating that the differentiation protocol is successful for all cell lines tested, with each iPSC differentiating into progenitors and ultimately DAN (**Figure S2**). The expression at 4- and 6-weeks of differentiation does not differ, indicating these markers cannot distinguish maturation levels. We found that raw absolute qPCR data was effectively processed by Auto-qPCR creating summary data, visualization and statistics for differential gene expression between conditions.

Relative quantification

In addition to Absolute quantification, the Auto-qPCR software also enables the processing of qPCR data obtained according to a relative quantification design. Contrary to absolute quantification, relative quantification does not require a calibration curve, and quantification (of transcripts) is based on the CT difference between a transcript of interest and one or more endogenous controls (**Figure 4A**). Relative qPCR is optimal for two kinds of comparisons: (1) detecting a difference in gene expression between two different conditions, and (2) detecting a difference between two transcripts within the same condition. Relative quantification can be expressed either as $RQ=2^{-\Delta CT}$, where samples are normalized to internal control(s), or $RQ=2^{-\Delta\Delta CT}$, where a given sample is considered as a calibrator for the unknown samples (**Figure 4B and 4C**).

To illustrate the functions of the program, we compared the expression levels of two different control cell lines at two developmental stages, indicated as D0 and D7, where D0 represents NPCs and D7 indicates 7 days of differentiation into cortical neurons. We measured the expression levels of the progenitor marker *PAX6* (paired box protein 6), two markers of neuronal differentiation, *GRIN1* (a subunit of the NMDA receptor) and *CAMK2A* (a subunit of calcium calmodulin kinase), and two housekeeping genes, *GAPDH* and *β-actin* as endogenous controls.

We used the Auto-qPCR app to process the same data twice, for a direct comparison of the two distinct relative quantification options. **Figure 4D** shows the mean expression from technical triplicates calculated by selecting the $RQ=2^{-\Delta CT}$ (indicated as delta CT in the web app). The ΔCT approach (not using a sample as calibrator) allows a comparison of the expression levels for the three different transcripts. We observed that relative to the endogenous controls, the D0 expression values for each transcript varied widely between the two cell lines tested. However, as expected for both cell lines, *PAX6* expression is higher at the D0 stage compared to D7. Conversely, both *GRIN1* and *CAMK2A* exhibited higher expression at the D7 stage compared to D0. Using the statistics module in the Auto-qPCR app, we compared the mean levels of each gene transcript at D0 and D7 using paired t-tests for each gene (**Figures 4E and 4F**). We found that although there were clear differences in expression, they were not significant between D0 and D7, likely a result of there only being two samples for each time point (**Table S9**). Interestingly, we found that the *CAMK2A* $RQ_{\Delta CT}$ was twice the level of *GRIN1* at D7 $RQ_{\Delta CT}$ (**Figure 4F**).

We next analysed this dataset with the $RQ_{\Delta\Delta CT}$ model (indicated as delta delta CT in the web app) where transcript levels are compared to both control gene expression (in this case *β-actin* and *GAPDH*) and a calibration sample; in this case we set one sample, AIW002-02-D0 arbitrarily as the reference sample (**Figure 4G**). Here we can easily compare expression in a test condition relative to a control condition by displaying the results as fold change in expression. All decreases are displayed as between 0 and 1 and all the increased expression levels are above 1 (**Figure 4C**). With the double normalization ($RQ_{\Delta\Delta CT}$), all values were expressed as a variation compared to the calibrator (AIW002-2-D0) as seen in **Figures 4G-I**. As in the $RQ_{\Delta CT}$ model, the

changes in gene expression from D0 to D7 were not significant (**Table S10**). Although the ratio of expression for a given gene in each cell line between DO and D7 remained unchanged, differential expression between genes can no longer be analysed. The $RQ_{\Delta\Delta CT}$ shown in **Figure 4H** showed that *PAX6* expression was higher at D0 than D7 and that *CAMK2a* and *GRIN1* expression were both higher at D7 than D0, as seen in **Figure 4E** using the $RQ\Delta_{CT}$ model. However, with the double normalization, the increase in *GRIN1* expression from D0 to D7 appears much larger than the increase in *CAMK2a* expression (**Figures 4H and I**), which was the opposite result from the single normalization model ($RQ_{\Delta CT}$) (**Figure 4E and 4F**). Our findings highlight the need to analyze data with attention to the biological question. Using only the $RQ_{\Delta\Delta CT}$ one might mistakenly believe the increase in *GRIN1* expression is greater than that of *CAMK2a.* With Auto-qPCR we provide a quick easy option to process the exported qPCR data with two different relative models. In our case, we found the same gene expression ratios between the two time points, but different expression gene levels using the different relative quantitation models.

Auto-qPCR can reprocess a published data set with the same results as manual processing

One of our objectives was to provide a tool for analyzing data from quantitative PCR experiments that had been generated with different machines. We took advantage of having the raw data from a set generated by the Gorwood lab, on a different machine (Opticon 2, Biorad). The original study measured gene expression in three sub cortical areas (subthalamic nucleus (STN), substantia nigra (SN) and globus pallidus (GP) of mice submitted to a place preference paradigm to cocaine (Kelai et al., 2008). Manual processing shows a significant increase in *Nrxn3* expression in the cocaine-treated group compared to control, specifically in the GP (**Figure 5A**).

We next used the raw data from this study (Kelai et al., 2008) to validate the Auto-qPCR software. We processed the data using the Auto-qPCR web app absolute quantification pipeline and normalized to *B2M*. The mean values of the technical replicates are shown for each biological replicate (mouse) **Figure 5B.** This summary data closely matched the manually calculated data (**Table S11)**. The standard method of removing outliers from technical replicates is to remove the replicate most different from the mean, if the CT standard deviation (std) is above 0.3. We have given the user the option in the Auto-qPCR software to adjust this threshold, termed 'cut-off' in the app. When data is processed manually, each CT-std value is analysed. In some cases, when the std value is close to 0.3, one replicate is clearly different from the other two, meaning that this divergent value will be removed. There are also instances in manual processing where no replicates are removed when the std is greater than 0.3, because the triplicate values are evenly distributed. Auto-qPCR has an option to account for this type of data when the user selects 'preserve highly variable values'. With this option a replicate is only removed if the median is far from the mean. We processed the *Nrxn3* expression data with a range of std cut-off values to display the difference in outcomes and with or without preserving

highly variable replicates (**Table S11**). We compared the variances generated by the differences between the dataset from manual treatment and the datasets collected (i) after application of a cut-off at 0.3, or (ii) after application of the same cut-off, with the possibility of preserving the outliers. We found that the preservation of highly variable value combined with a cut-off at 0.3 generate a 20% decrease in the variance between manual and automatic treatments **(Table S12)**. With this treatment, the software also preserved a value that was falsely estimated as an outlier by manual processing, which illustrates the subjectivity of the user with respect to the decision to retain or exclude a value based on criteria of divergence, especially when working with low numbers. Together, our analysis suggests that applying two rules of data filtering provides a more systematic data analysis method and minimizes interindividual bias. Here we applied the standard cut-off of 0.3 and preserved highly variable replicates, appropriate for the highly variable and RNA level experimental samples we are analyzing.

Auto-qPCR also permits statistical groups to be designated in the sample name or in a specific group column, which can be added into the qPCR data during the plate set up. To allow for statistical analysis of this data, we added a grouping columns into the raw data files: 'Treatment' (Control, Cocaine), 'Region' (STN,GP,SN) and both together, 'T_R' (STN_Control, STN_Cocaine, GP_Control, GP_Cocaine, SN_Control, SN_Cocaine) and using the Auto-qPCR statistics module, we reanalysed the effect of drug treatment and brain regions on expression of *Nrxn3* across several parameters. We first compared the overall effect of cocaine on expression after pooling the three brain regions and found that although the expression of *Nrxn3* was increased across brain regions with cocaine treatment, there was no overall significant effect of drug treatment (**Figure 5C and Table S13)**. Comparing the three brain regions while pooling together control and cocaine treatment showed a significant difference in expression across brain regions. Post-hoc analysis revealed *Nrxn3* expression in the STN was significantly lower than in the GP and SN (**Figure 5D and Table S14).** When we considered each brain region with and without treatment as independent conditions, and individual mice as biological replicates and used a one way ANOVA followed by posthoc tests using multiple t-test with a correction for multiple comparisons we find cocaine significantly increased *Nrxn3* expression specifically in the GP and not in the SN or STN (**Figure 5E and Table S15**). This was in agreement with the previous manual analysis (Kelai et al., 2008). To apply the identical statistical treatment as originally presented, we performed a two-way ANOVA followed by a repeated measures t-tests with FDR correction on the interaction variable between treatment and brain region, using Auto-qPCR and found the same results as the one-way ANOVA (**Figure 5F and Table S16**) and a t-test of the GP alone (**Figure 5G**). Together the data shows that the Auto-qPCR software is capable of processing data generated by another machine and the results match those processed manually.

Discussion

Auto-qPCR – a web app for Q-PCR data analysis and visualization

This paper presents a new software for qPCR analysis, and we provide examples of the functionalities for the web app and how it can be applied with qPCR experimental datasets generated from DNA (genomic instability assay) or cDNA amplification of RNA transcripts (absolute and relative quantification data). We have also summarized the computational bases of relative and absolute quantifications performed by Auto-qPCR, which is important for users to understand when selecting an experimental design. Additional functionalities included with the Auto-qPCR web app is a statistical module that will (1) be applicable to the majority of qPCR analysis experiments, and (2) provide a correction across multiple tests to mitigate against false positives. As not all experimental designs require differential analyses, we also provide the user with a choice of statistical analyses or simply calculating normalized RNA concentrations. Furthermore, the web app can be used with no installation or login requirements. We have created an easy to use program that is completely free and open source, able to process data from different qPCR machines and all common experimental designs, beneficial to any lab performing qPCR experiments.

A comparison of Auto-qPCR relative to available to qPCR analysis software

Several steps are required for qPCR experiments, from the design to the presentation of the differential expression analysis. We have created Auto-qPCR to process all major types of qPCR designs. Given the importance of qPCR in molecular biology, other programs are available to perform many steps of the qPCR data treatment (Pabinger et al., 2009; Pabinger et al., 2014; Rancurel et al., 2019; Zanardi et al., 2019; Krahenbuhl et al., 2020). The Q-PCR and PIPE-T programs were designed to treat and display qPCR data generated according to a relative quantification model (Pabinger et al., 2009; Zanardi et al., 2019) . SATQPCR is a web app that treats qPCR data using the relative quantification model and performs differential analyses. However, it does not take the exported csv qPCR data and requires preformatted data that has already been manually manipulated in txt file format (Rancurel et al., 2019). Finally, ELIMU-MDx, is a web-based interface conceived to collect specific information regarding qPCR assays for diagnostic purposes. EILMU-MDx functions as a data management system, processes qPCR data generated using the absolute quantification method and requires an account and login information (Krahenbuhl et al., 2020).

Reviewing different software published to serve similar purposes as ours highlights the unique characteristics of Auto qPCR, as no other web app combines all the features we have included in our software. First as a web app, Auto-qPCR does not require installation or a user login and can be accessed from any device connected to internet. We also provide the option for users to install the program onto their computer if they

want to work on their analysis off-line. Second, data processed by Auto-qPCR does not require any preformatted file to be generated manually. Instead, once the qPCR experiment is complete, our program takes the csv export file directly from the thermocycler so there is no copy/paste or formatting step to be done by the user. Third, Auto-qPCR can manage the data from a single or from multiple separate absolute files at once, as well as batch process multiple csv files from a relative quantification. The program creates a clean data set and summary data table. Fourth, unlike the other software mentioned above, Auto-qPCR includes three different models, conceived to support qPCR data generated from absolute and two methods of relative quantification designs. No other program provides the option of choosing between the two relative quantification methods. Fifth, we provide normalization to multiple reference genes and calculate the mean normalized value for each replicate, and not the sample mean, an important feature implemented in relatively few other programs. This avoids the RNA quantity value being influenced by extreme values. Sixth, we extend the use of the program to suit qPCR data from DNA quantification. Finally, we provide an extensive statistics module for calculating differential gene expression that requires no additional input files. Options are included for experimental designs that include two or more sample comparisons (t-test, one- and two-way ANOVA and the equivalent non-parametric tests) and automatically generates bar charts for data visualization. We have created a unique, easy to use qPCR analysis program that can benefit any researcher or lab that needs to analyze qPCR data on a regular basis, by saving time, avoiding errors and generating reproducible, figure-ready plots.

Calculating gene expression using two "relative quantification" methods

Auto-qPCR provides users the option for relative quantification by two methods: expression relative to endogenous control genes only (ΔCT method) or relative to endogenous genes and also normalized to a control condition (ΔΔCT method). Although the ΔΔCT method is considered the gold standard to express, in one number, the variation in gene expression between two conditions and the amplitude of that change in expression (Schmittgen and Livak, 2008), it does not account for inter gene expression variation within the control condition (Yuan et al., 2006). The differences between quantifying relative expression with or without a control condition used as a calibrator, are clearly demonstrated above (**Figure 4**). Expression levels of *GRIN1* and *CAMK2α* calculated with either relative quantification model were increased at seven days of differentiation (D7) compared to day zero (DO). However, we also found that *GRIN1* and *CAMK2A* had different levels in the baseline condition (ΔCT), thus we observe that information is lost when using a ΔΔCT normalization. For relative quantification using a ΔΔCT normalization we measured a fold change of variation compared to a control condition for a given gene (Rao et al., 2013), but information about differences of expression between two genes in control condition were not observed (**Figure 4F**). We have provided both

the gold standard method of relative quantification and a method to calculate gene expression without a reference sample, to allow users to quickly determine expression changes without loosing information about the level of expression levels in control conditions.

Improving efficiency of data processing and reproducibility of analyzes between users

Reprocessing the external dataset highlighted two main advantages of treating qPCR dataset with a program. First, manual analysis of qPCR data is time consuming. Second, comparing both data treatments (manual and program-assisted) has shown that one important source of variation between results of manual analysis is the inconsistent rules used for data exclusion. Although, removing one outlier from technical replicates, in the vast majority of cases, improves the CT standard deviation (std) by decreasing it under the commonly accepted threshold of 0.3; in many cases researchers decide to keep a technical replicate even if the CT-std value is above 0.3. These judgement calls frequently occur when transcripts have low expression levels and the high variance between technical replicates does not permit a decision based on the adjustment of the CT std. To account for these situations, we incorporated a second rule for data inclusion/exclusion based on the distance between the arithmetic mean and the median value of technical replicates to determine the most acceptable set of technical replicates. Applying such an algorithm to the user's judgement removes variability and potential bias in the resulting normalized gene expression levels. We were able to reprocess external data using Auto-qPCR and acquired the same summary output, reaching the same conclusions as the initial study. We showed that Auto-qPCR can process data from different PCR machines and matched the expected outcome from manual processing without the risk of bias or errors. Using a double rule for data inclusion/exclusion for highly variable signal between technical replicates, the program provides a unique treatment that will considerably reduce the risk of variability and mistakes generated by and between users during manual data processing.

Additional uses of the Auto-qPCR software

The Auto-qPCR program has many other potential uses not including in this manuscript, and one such use is for analyzing data from a chromatin immunoprecipitation experiment followed by specific DNA amplification. There are several ways of normalizing outputs of chromatin immunoprecipitation ChiP assays. DNA amplification of a region of interest in an immunoprecipitated sample can either be compared to the amplification of non-immunoprecipitated fraction (Nagaki et al., 2003); or a fold enrichment of DNA amplification estimated by comparing the amplification level of a candidate region for protein binding to a DNA region that is known to be unbound (Mathieu et al., 2005). In a previous study we chose the second approach (Maussion et al., 2015), however thedata could be treated using Auto-qPCR and analyzed using

either the absolute or the relative quantification models. The absolute quantification method would permit testing the primer efficiency through the calibration curve (Brankatschk et al., 2012), and the DNA target amplification would be normalized to an unbound DNA. Alternatively, the level of DNA/protein interaction can be estimated using the relative quantification models. One or several regions, known to be unbound by a protein of interest, can be defined as endogenous control(s). DNA/protein interactions can be quantified and compared using either ΔCT or ΔΔCT delta CT methods depending on whether the experimental design has a reference condition.

By providing the absolute and two relative models to process qPCR data, Auto-qPCR is flexible enough to let the user choosing the most appropriate model to use, based on the information available on the DNA regions to amplify and analyze.

Standardizing and automating data processing allows for better experimental design

The Auto-qPCR program was conceived to treat, analyze and display qPCR data generated using either relative or absolute quantification designs, while limiting errors related to manual processing. The absolute and relative quantification procedures provides distinct information about gene expression and DNA integrity. Absolute quantification is accurate when the biological question is examining changes in individual gene expression over various experimental conditions. The relative model without a reference sample can detect differences in expression levels between genes within a given condition. When the variation of gene expression between two or more conditions is addressed, the relative quantification model using a reference sample is suitable. Data processing tools will never replace or supplement appropriate experimental design and statistical power. The conditions included with the design and interpretation of the results still remain in the user's hand. We have provided a tool that will provide easy, reproducible analysis without user errors for unlimited samples. Although, we cannot computationally remove the need for replication and controls, analysis time will no longer be a limitation. Auto-qPCR can also assist the user in larger experimental designs. The elements required to support qPCR data generated in duplex (from two probes) data across several plates using relative quantification designs are already present in the program. All these possibilities were included to allow the user to focus on determining the most accurate experimental design to answer biological questions.

Author Contributions

GM and RAT have conceptualized the program. ID, GG, EC and RAT have written and tested the program. RAT managed the program development and GitHub repository. RAT ran all the analysis using the webapp. GG built the graphical user interface and website. TJPS has transferred the website to run online through a virtual machine. GM generated the qPCR data used to test the absolute and relative quantification models of Auto-qPCR program. CXQC, NA and ANJ, have extracted DNA and performed the PCR used to improve the pipeline related to genomic instability model of Auto-qPCR program. SK, NR and PG have generated the data used as the external data set for the figure 5 of the manuscript. RAT, ID and GM made the figures. GM, RAT, LKB and TMD wrote the manuscript.

Acknowledgements:

Thanks to Ivan Castanon Niconoff for helping create and set up the virtual machine used to host the Auto-qPCR web app. Thanks to the members of the Early Drug Discovery Units, especially Maria José Castellanos Montiel, Vincent Soubannier and Nguyen-Vi Mohamed, who tested the web app during its final beta testing phases. Special thanks to Genevieve Dorval for helping with support for the project.

Funding

TMD funding to support this project through the McGill Healthy Brains for Healthy Lives (HBHL) initiative, the CQDM FACs program, the Alain and Sandra Bouchard Foundation, the Ellen Foundation and the Mowfaghian Foundation. TMD is supported by a project grant from CIHR (PJT – 169095).

RAT was funded by a Healthy Brains for Healthy Lives Fellowship.

References

Artyukhov, A.S., Dashinimaev, E.B., Tsvetkov, V.O., Bolshakov, A.P., Konovalova, E.V., Kolbaev, S.N., et al. (2017). New genes for accurate normalization of qRT-PCR results in study of iPS and iPS-derived cells. *Gene* 626**,** 234-240. doi: 10.1016/j.gene.2017.05.045.

Beggs, A.H., Koenig, M., Boyce, F.M., and Kunkel, L.M. (1990). Detection of 98% of DMD/BMD gene deletions by polymerase chain reaction. *Hum Genet* 86(1)**,** 45-48.

Bell, S., Peng, H., Crapper, L., Kolobova, I., Maussion, G., Vasuta, C., et al. (2017). A Rapid Pipeline to Model Rare Neurodevelopmental Disorders with Simultaneous CRISPR/Cas9 Gene Editing. *Stem Cells Transl Med* 6(3)**,** 886-896. doi: 10.1002/sctm.16-0158.

Brankatschk, R., Bodenhausen, N., Zeyer, J., and Burgmann, H. (2012). Simple absolute quantification method correcting for quantitative PCR efficiency variations for microbial community samples. *Appl Environ Microbiol* 78(12)**,** 4481-4489. doi: 10.1128/AEM.07878-11.

Bustin, S.A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol* 25(2)**,** 169-193.

Charbonnier, F., Raux, G., Wang, Q., Drouot, N., Cordier, F., Limacher, J.M., et al. (2000). Detection of exon deletions and duplications of the mismatch repair genes in hereditary nonpolyposis colorectal cancer families using multiplex polymerase chain reaction of short fluorescent fragments. *Cancer Res* 60(11)**,** 2760-2763.

Chen, C.X.Q., Abdian, N., Maussion, G., Thomas, R.A., Demirova, I., Cai, E., et al. (2021). Standardized quality control workflow to evaluate the reproducibility and differentiation potential of human iPSCs into neurons. *bioRxiv***,** 2021.2001.2013.426620. doi: 10.1101/2021.01.13.426620.

Chen, E.S., Lauinger, N., Rocha, C., Rao, T., and Durcan, T.M. (2019a). Generation of dopaminergic or cortical neurons from neuronal progenitors. *Zenodo*. doi: 10.5281/zenodo.3361005.

Chen, E.S., Rocha, C., Loignon, M., Peng, H., Rao, T., and Durcan, T.M. (2019b). Induction of Dopaminergic or Cortical neuronal progenitors from iPSCs *Zenodo*. doi: 10.5281/zenodo.3364831.

D'Haene, B., Vandesompele, J., and Hellemans, J. (2010). Accurate and objective copy number profiling using real-time quantitative PCR. *Methods* 50(4)**,** 262-270. doi: 10.1016/j.ymeth.2009.12.007.

Dahl, J.A., and Collas, P. (2007). Q2ChIP, a quick and quantitative chromatin immunoprecipitation assay, unravels epigenetic dynamics of developmentally regulated genes in human carcinoma cells. *Stem Cells* 25(4)**,** 1037-1046. doi: 10.1634/stemcells.2006-0430.

De la Vega, F.M., Lazaruk, K.D., Rhodes, M.D., and Wenz, M.H. (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System. *Mutat Res* 573(1-2)**,** 111-135. doi: 10.1016/j.mrfmmm.2005.01.008.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464(7291)**,** 1071-1076. doi: 10.1038/nature08975.

Kelai, S., Maussion, G., Noble, F., Boni, C., Ramoz, N., Moalic, J.M., et al. (2008). Nrxn3 upregulation in the globus pallidus of mice developing cocaine addiction. *Neuroreport* 19(7)**,** 751-755. doi: 10.1097/WNR.0b013e3282fda231.

Krahenbuhl, S., Studer, F., Guirou, E., Deal, A., Machler, P., Hosch, S., et al. (2020). ELIMU-MDx: a web-based, open-source platform for storage, management and analysis of diagnostic qPCR data. *Biotechniques* 68(1), 22-27. doi: 10.2144/btn-2019-0064.

Kriks, S., Shim, J.W., Piao, J., Ganat, Y.M., Wakeman, D.R., Xie, Z., et al. (2011). Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature* 480(7378), 547-551. doi: 10.1038/nature10648.

Magnuson, V.L., Ally, D.S., Nylund, S.J., Karanjawala, Z.E., Rayman, J.B., Knapp, J.I., et al. (1996). Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotechniques* 21(4), 700-709. doi: 10.2144/96214rr03.

Mathieu, O., Probst, A.V., and Paszkowski, J. (2005). Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. *EMBO J* 24(15), 2783-2791. doi: 10.1038/sj.emboj.7600743.

Maussion, G., Diallo, A.B., Gigek, C.O., Chen, E.S., Crapper, L., Theroux, J.F., et al. (2015). Investigation of genes important in neurodevelopment disorders in adult human brain. *Hum Genet* 134(10), 1037-1053. doi: 10.1007/s00439-015-1584-z.

Milne, T.A., Zhao, K., and Hess, J.L. (2009). Chromatin immunoprecipitation (ChIP) for analysis of histone modifications and chromatin-associated proteins. *Methods Mol Biol* 538, 409-423. doi: 10.1007/978-1-59745-418-6_21.

Mukhopadhyay, A., Deplancke, B., Walhout, A.J., and Tissenbaum, H.A. (2008). Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in Caenorhabditis elegans. *Nat Protoc* 3(4), 698-709. doi: 10.1038/nprot.2008.38.

Mullis, K.B., and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol* 155, 335-350.

Nagaki, K., Talbert, P.B., Zhong, C.X., Dawe, R.K., Henikoff, S., and Jiang, J. (2003). Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. *Genetics* 163(3), 1221-1225.

Ovstebo, R., Haug, K.B., Lande, K., and Kierulf, P. (2003). PCR-based calibration curves for studies of quantitative gene expression in human monocytes: development and evaluation. *Clin Chem* 49(3), 425-432. doi: 10.1373/49.3.425.

Pabinger, S., Rodiger, S., Kriegner, A., Vierlinger, K., and Weinhausel, A. (2014). A survey of tools for the analysis of quantitative PCR (qPCR) data. *Biomol Detect Quantif* 1(1), 23-33. doi: 10.1016/j.bdq.2014.08.002.

Pabinger, S., Thallinger, G.G., Snajder, R., Eichhorn, H., Rader, R., and Trajanoski, Z. (2009). QPCR: Application for real-time PCR data management and analysis. *BMC Bioinformatics* 10, 268. doi: 10.1186/1471-2105-10-268.

Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 29(9), e45. doi: 10.1093/nar/29.9.e45.

Rancurel, C., van Tran, T., Elie, C., and Hilliou, F. (2019). SATQPCR: Website for statistical analysis of real-time quantitative PCR data. *Mol Cell Probes* 46, 101418. doi: 10.1016/j.mcp.2019.07.001.

Rao, X., Huang, X., Zhou, Z., and Lin, X. (2013). An improvement of the 2^(-delta delta CT) method for quantitative real-time polymerase chain reaction data analysis. *Biostat Bioinforma Biomath* 3(3), 71-85.

Saiki, R.K., Bugawan, T.L., Horn, G.T., Mullis, K.B., and Erlich, H.A. (1986). Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature* 324(6093), 163-166. doi: 10.1038/324163a0.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., et al. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230(4732), 1350-1354. doi: 10.1126/science.2999980.

Scharf, S.J., Horn, G.T., and Erlich, H.A. (1986). Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science* 233(4768), 1076-1078. doi: 10.1126/science.3461561.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc* 3(6), 1101-1108. doi: 10.1038/nprot.2008.73.

Shi, R., and Chiang, V.L. (2005). Facile means for quantifying microRNA expression by real-time PCR. *Biotechniques* 39(4), 519-525. doi: 10.2144/000112010.

Tosca, L., Feraud, O., Magniez, A., Bas, C., Griscelli, F., Bennaceur-Griscelli, A., et al. (2015). Genomic instability of human embryonic stem cell lines using different passaging culture methods. *Mol Cytogenet* 8, 30. doi: 10.1186/s13039-015-0133-8.

Varkonyi-Gasic, E., Wu, R., Wood, M., Walton, E.F., and Hellens, R.P. (2007). Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods* 3, 12. doi: 10.1186/1746-4811-3-12.

Wong, M.L., and Medrano, J.F. (2005). Real-time PCR for mRNA quantitation. *Biotechniques* 39(1), 75-85. doi: 10.2144/05391RV01.

Ye, S., Dhillon, S., Ke, X., Collins, A.R., and Day, I.N. (2001). An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res* 29(17), E88-88. doi: 10.1093/nar/29.17.e88.

Yoshihara, M., Hayashizaki, Y., and Murakawa, Y. (2017). Genomic Instability of iPSCs: Challenges Towards Their Clinical Applications. *Stem Cell Rev Rep* 13(1), 7-16. doi: 10.1007/s12015-016-9680-6.

Yuan, J.S., Reed, A., Chen, F., and Stewart, C.N., Jr. (2006). Statistical analysis of real-time PCR data. *BMC Bioinformatics* 7, 85. doi: 10.1186/1471-2105-7-85.

Zanardi, N., Morini, M., Tangaro, M.A., Zambelli, F., Bosco, M.C., Varesio, L., et al. (2019). PIPE-T: a new Galaxy tool for the analysis of RT-qPCR expression data. *Sci Rep* 9(1), 17550. doi: 10.1038/s41598-019-53155-9.

Figure Legends

Figure 1: Workflow of a qPCR experiment

Schematic representation of common qPCR assays: genomic stability assay to detect DNA deletions or duplication events (green line), two methods to quantify RNA (cDNA) using either absolute (red line) or relative quantification designs (blue lines). qPCR experiments can be sub divided in two parts: the sample preparation and running the PCR machine (Experimental Work-Flow) and the data analyses (Auto QPCR Program). The preparation of the experiment includes nucleic acid extraction followed by a cDNA synthesis step (for RNA) and the *in silico* design of the PCR plate layout. Nucleic acid preparations must be accurately diluted. For the absolute model, a standard curve must be created. The experimental design of the PCR plate, including the chemistry (fluorophore, primer mix), the status of the samples, and the transcripts or DNA region that are going to be amplified, must be generated *in silico.* After having defined the parameters of the qPCR reactions (number of PCR cycles and length of the different steps (denaturation, hybridization and elongation), and the temperatures), the PCR is run. The exported data from the thermocycler, converted to csv, is entered into the Auto-qPCR software and the model matching the experimental design and parameters for analysis are selected. The software will reformat the data, quantify each sample normalized to controls, and create spreadsheets and graphs to visualize the data analyses, all of which will be included in a zip file for the user to save.

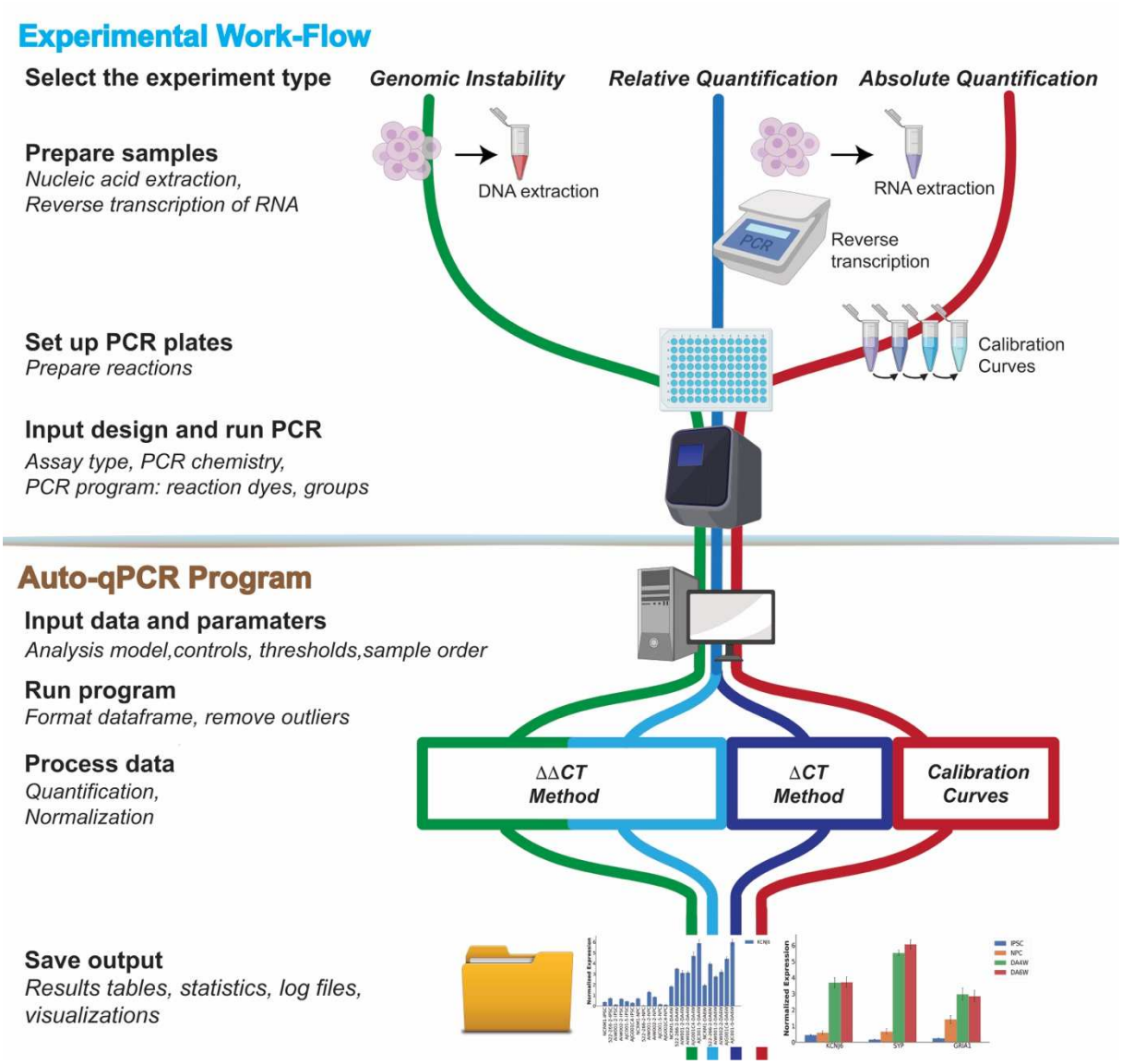
Figure 2: Auto-qPCR can process PCR genomic stability data. A) Screen capture of the Auto-qPCR web-app. **B)** Simplified schematic of PCR workflow showing the analysis for genomic instability in green. The DNA copy number is quantified with the same formula as the delta-delta CT relative quantification model. **C)** The calculations carried out for genomic instability testing (delta-delta CT). Top, the general formula used where the CT values for each chromosome were normalized to a region of interest and then to a reference sample. Middle, the reference DNA region (CHR4) and the reference sample (Normal) used in this dataset. Bottom, the confidence interval for determining a genomic instability, insertion or deletion event. **D)** Bar chart showing the output from Auto-qPCR program running the genomic instability model. Four different iPSC cell lines are indicated and compared to the control sample. Normalized signals for all four cell lines are in the confidence interval defined by the control sample.

Figure 3: Auto-qPCR can process quantitative qPCR data using a standard curve to perform statistical analysis. Output of Auto-qPCR processing using the absolute model. **A)** Illustration of a calibration curve displaying 8 serial dilution points of four-time dilution which covers cDNA quantities from 0.003053 to 50 ng

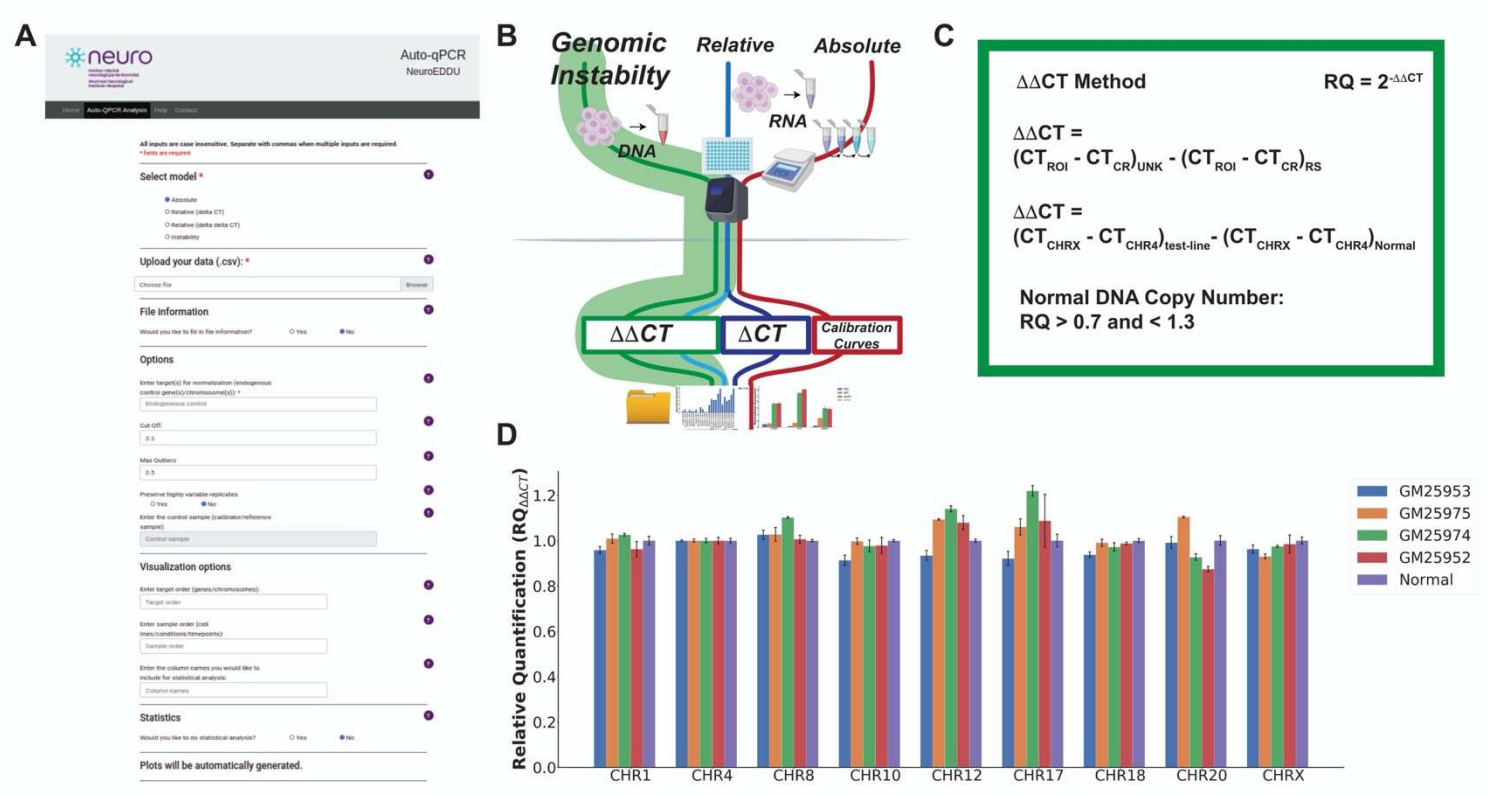
and establishes the linear relationship between CT values (y-axis) and the log2[RNA]. **B)** Schematic of PCR workflow showing the pipeline for the absolute quantification using a standard curve in red. **C)** Formula used to process a real-time PCR experiment using an absolute quantification design. Top, general formula where the linear relation between the logarithm of RNA concentration and the CT value is provided by the calibration curve. The normalized quantification is expressed as a ratio between concentrations for the gene of interest and the endogenous control(s) estimated from their respective calibration curves. Bottom, the variables specific to this dataset are shown in the general formula. **D)** Bar chart showing the output from Auto-qPCR program using the absolute model for the normalized expression of the gene *KCNJ6* for 6 cell lines at 4 different developmental stages (iPSC: induced pluripotent stem cells; NPC, Neural progenitor cells; DA4W, dopaminergic neurons at 4 weeks, DA6W: Dopaminergic neurons at 6 weeks). **E)** and **G)** Bar charts showing the average expression levels obtained from the three technical replicates for each cell line and time point for the three genes (S*YP, KCNJ6* and *GRIA1),* normalized with two housekeeping genes (*ACTB*: *beta-actin*, *GAPDH*). **E)** Mean RNA expression grouped by genes on the x-axis, cell lines and time points are indicated in legend. **G)** Mean RNA expression grouped by cell lines and time points; the gene transcripts quantified are indicated in the legend. **F)** and **H)** Bar charts showing the mean expression levels of S*YP, KCNJ6* and *GRIA1* for four developmental stages (n=6 cell lines). **F)** Grouped by genes (x-axis), time points are indicated in the legend. **H)** Grouped by time points (x-axis), the genes are indicated in the legend. One way ANOVAs across differentiation stages for *KCNJ6*, *SYP* and *GRIA1* ($p < 0.001$, $p < 0.001$, $p = 0.002$).

Figure 4: Auto-qPCR can process quantitative PCR data using two different relative models. Output of Auto-qPCR using the relative quantification with both the ΔCT and ΔΔCT models. **A)** Amplification curves illustrating a difference of cycle threshold values (ΔCT) between a gene of interest and an endogenous control. **B)** Schematic of PCR workflow showing the two methods to calculate relative RNA quantity, ΔCT in dark blue and ΔΔCT in light blue. **C)** Formula used to perform a qPCR using relative quantification models, according the ΔCT (right), or the ΔΔCT methods (left). **D-F)** Bar charts showing the output of the delta-CT model ($RQ^{\Delta CT}$). **G-I**) Bar charts showing the output from the delta-delta-CT model ($RQ^{\Delta\Delta CT}$). **D) and G)** Mean normalized gene expression values from technical replicates for the genes *PAX6*, *CAMK2A* and *GRIN1* indicated on the x-axis for 2 cell lines at two stages of differentiation (D0: Neural progenitor cells, and D7: cortical neurons at 7 days of differentiation) as indicated. **E)** and **H)** Statistics output showing the mean gene expression from two cell lines at two stages of differentiation indicated, for the three genes indicated on the x-axis. F) and I) Statistics output showing the mean expression values for two cell lines at two time points on the x-axis and the three genes indicated. Differential expression between D0 and D7 is not significant (*PAX6* $p = 0.40$, *CAMK2A* $p=0.18$, *GRIN1* $p=0.16$), t-tests, n=2.

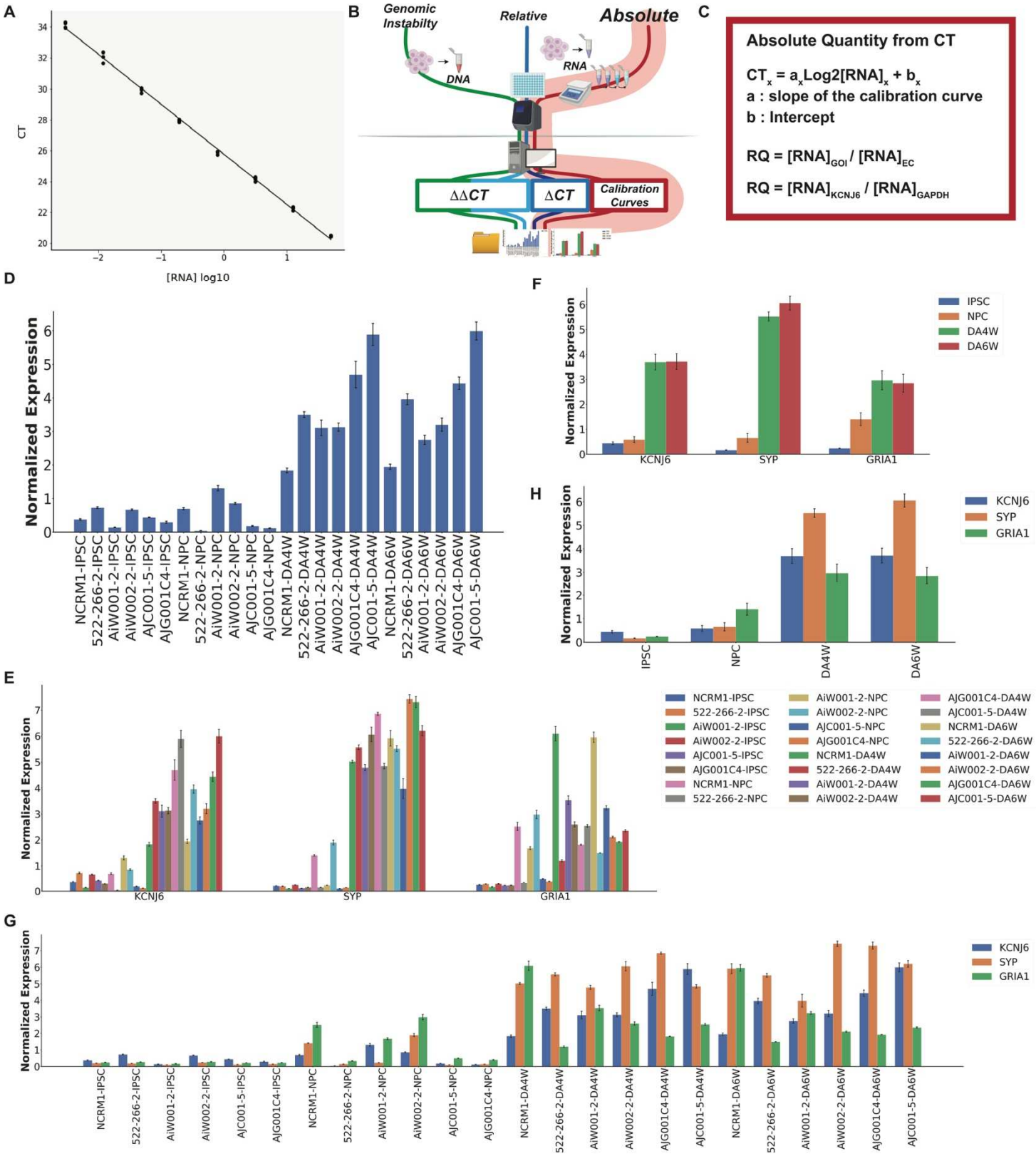
Figure 5: Auto-qPCR can process data from different thermocyclers and produce the same results as manual processing. A) Bar chart showing the mean *Nrxn3* expression level normalized to *B2M levels assessed with an absolute quantification design manually processed and plotted in Prism, grouped by brain regions (STN: subthalamic nucleus, GP: globus paladus*, SN: substantia nigra) on the x-axis, with and without cocaine treatment. **B)** Output of Auto-qPCR processing the same dataset. *Nrxn3* normalized expression levels from technical replicates for each biological sample. The treatment conditions are indicated below the x-axis. **C)** Statistics output of Auto-qPCR program comparing cocaine and control groups. *Nrxn3* normalized expression levels in the combined brain regions. Expression is not significantly different, p=0.113, t-test, n=13. **D)** Auto-qPCR statistical output showing mean *Nrxn3* expression combining treatments and comparing the three brain regions. One way ANOVA shows significant effect of brain regions, FDR adjusted $p < 0.001$, n=9 for GP and SN, n= 10 STN. **E)** Bar chart of Nrxn3 expression shown as six groups distinguished by brain region and treatment generated by Auto-qPCR program after a one-way ANOVA, $p < 0.001$, n=4 or 5. Posthoc analysis using multiple t-test with FDR correction comparing treatment at each brain region: SNT p=0.990, GP p=0.033 , SN p=0.413. **F)** Bar chart of *Nrxn3* average normalized by brain region (x-axis) and treatment, generated by Auto-qPCR program after a two-way ANOVA, brain region $p < 0.001$, treatment p= 0.2265, n=4 or 5.Posthoc analysis using multiple t-test with FDR correction comparing each brain region with and without cocaine: SNT p=0.0.998, GP p=0.053 and p-unadjusted = 0.017, SN p=0.619 **G)** Bar chart of the average *Nrxn3* normalized expression levels in the GP compared between the two groups with a t-test (p = 0.0176).



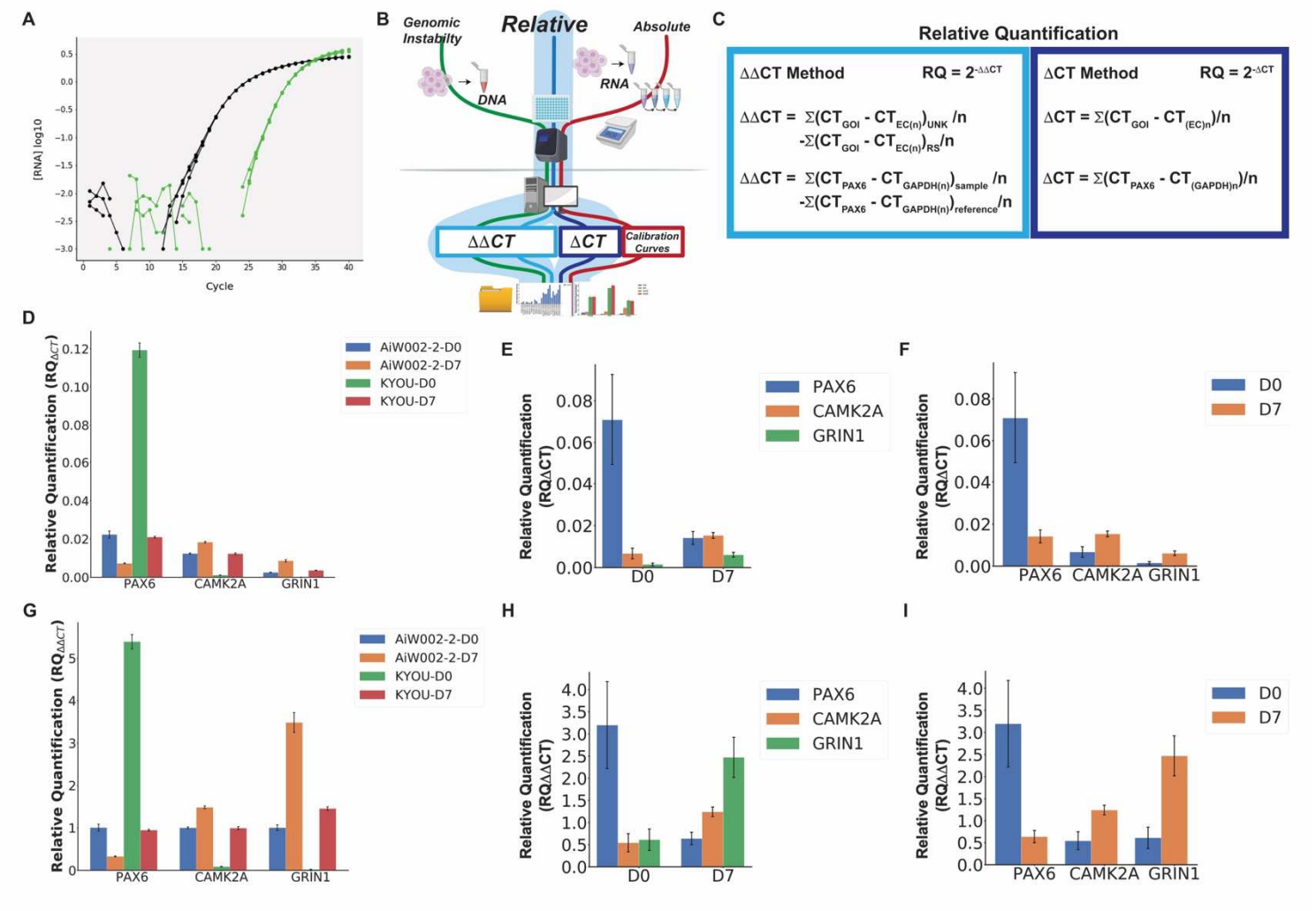
Maussion, Thomas et al. Figure 1



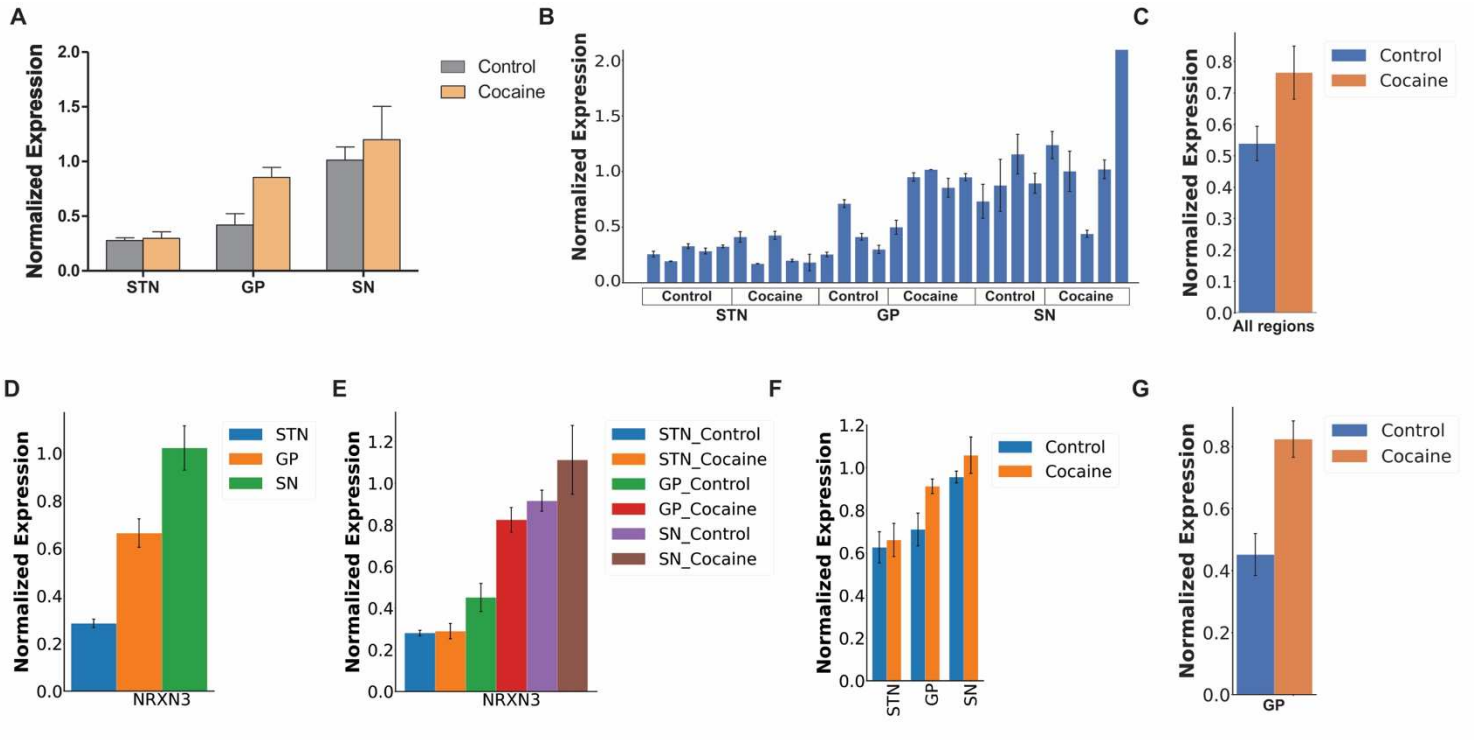
Maussion, Thomas et al. Figure 2



Maussion, Thomas et al. Figure 3



Maussion, Thomas et al. Figure 4



Maussion, Thomas et al. Figure 5

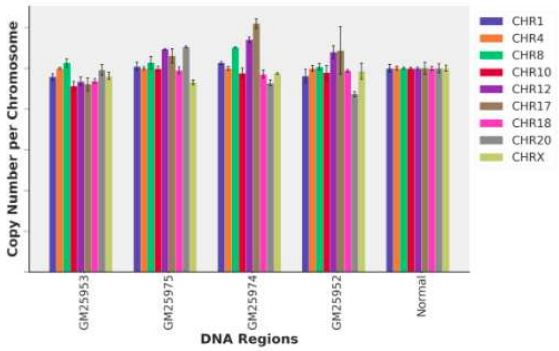
Figure S1: Example output from Auto-qPCR using the genomic instability model. A) The Log.txt output from the file generated by Auto-qPCR. The file lists the steps completed by the program and the inputs from the web interface. This example is from the genomic instability analysis. The selection for statistical analysis is also shown in the text file. Using the log file, the exact analysis can be repeated because all the settings are recorded. **B**) Bar chart showing an alternative visualization for the genomic instability assay where the data is grouped by cell lines on the x-axis and colours indicated in the legend represent the regions of chromosomes tested.

Figure S2: Example images of AJG001-C4 at four stages of development (iPSCs, NPCs, as well as 4 and 6 week DANs). A) iPSCs stained for pluripotency markers (Nanog, Tra1-60, SSEA4, OCT3-4 as indicated), together with Hoechst and shown as merged images on the right. **B**) Neural precursor cells (NPCs) expressing dopaminergic lineage (SOX1 and OTX2), proliferation (Ki67) and neural progenitors (Nestin) markers. **C**) Dopaminergic neurons after 4 and 6 weeks of differentiation stained with neuronal marker Tuj1 in all images and dopaminergic markers FOXA2, GIRK2 and TH as indicated.

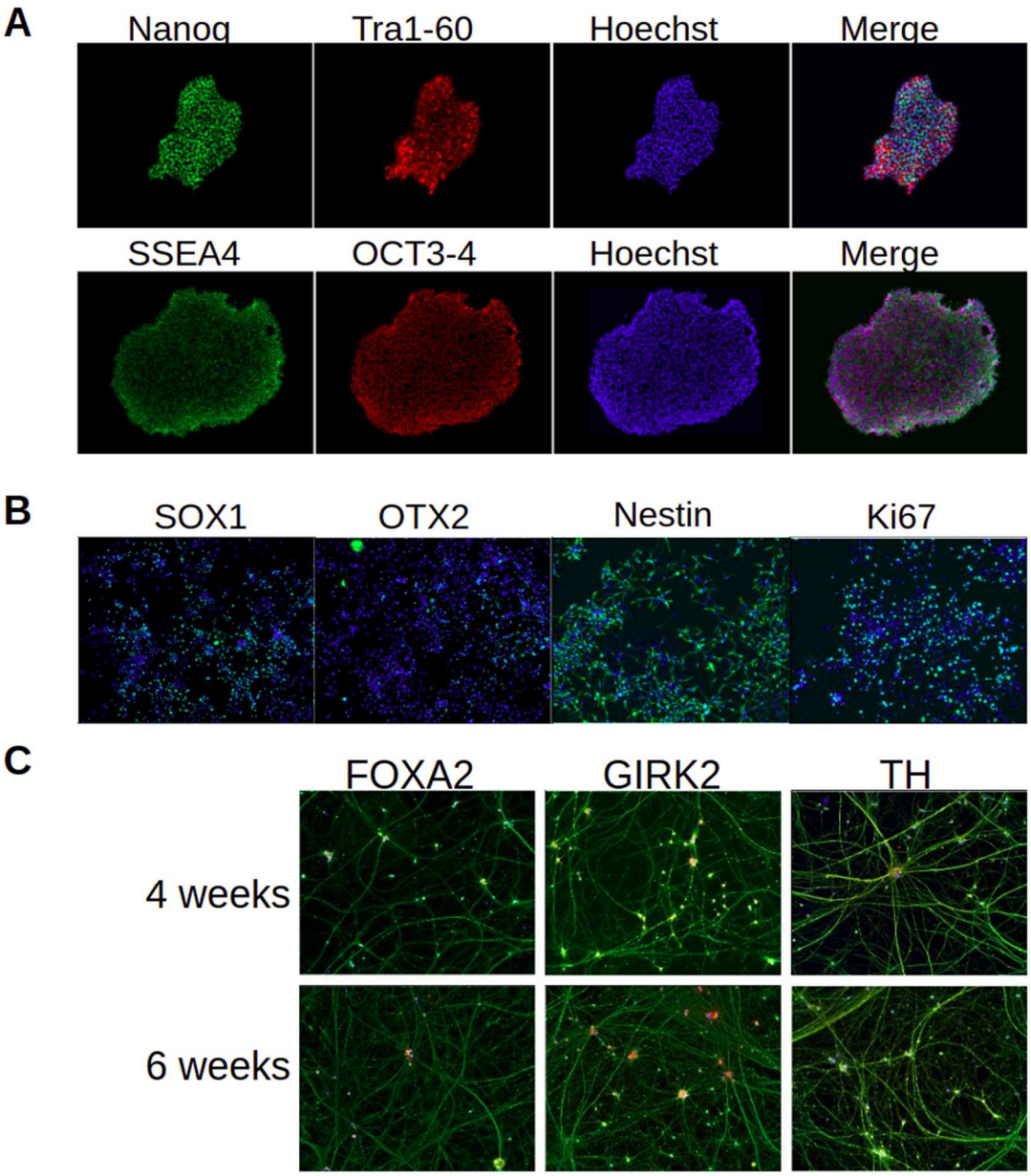
A

```
Started
Files upload complete.
Gene names if they are included in file names:
Model: instability
Quencher:
Task: UNKNOWN
Endogenous control genes: CHR4
Cut-off: 0.3
Maximum Outliers: 0.5
Target Order: CHR1,CHR4,CHR8,CHR10,CHR12,CHR17,CHR18,CHR20,CHRX
Sample Order: GM25953,GM25975,GM25974,GM25952,Normal
Control Sample: Normal
Additional column names:
Number of groups: None
Group column name:
Group name:
Column name A:
Column Name B:
Group names for column A:
Group names for column B:
Repeated measures: False
Normal distribution: True
Clean data and summary data are created
Plots of the summary data are created.
```

B



Maussion, Thomas et al. Supplemental Figure S1



Maussion, Thomas et al. Supplemental Figure S2

Supplemental Tables

Table S1: Overview of cell lines: Human-derived induced pluripotent stem cells used.

Cell line	Donor Age	Sex	Cell Type	Reprogramming Method
GM25952	10	F	Fibroblast	Episomal
GM25953	43	F	Fibroblast	Episomal
GM25974	7	F	Fibroblast	Episomal
GM25975	37	F	Fibroblast	Episomal
522-2666-2	NA	NA	Lymphocytes	Retrovirus
AIW001-2	48	F	PBMCs	Retrovirus
AIW002-2	37	M	PBMCs	Retrovirus
NCRM1	NA	M	Cord Blood	Episomal
AJG001-C4	37	M	PBMCs	Episomal
AJC001-5	37	M	Fibroblast	Retrovirus
KYOU-DRX0190B	36	F	Fibroblast	Retrovirus

Table S2: Taqman primers/probe sets. The primer/probe sets listed were used to generate the data presented in Figures 3 and 4 and test the absolute and relative quantification models to assess gene expression levels by Auto-qPCR web app. The primer/probe sets were selected from the assays available on the Thermo Fisher Scientific web site and chosen to cover the most important number of alternative transcripts for a given gene. With the exception of the assay for GAPDH, the amplicons overlap two exons, avoiding amplification of genomic DNA that could remain from incomplete DNAse digestion. The refseq sequence used for designing the primer/probe set assay is shown.

Gene Symbol	Gene Name	Location	Assay Reference	Exon Boundaries	Reference Accession
ACTB	Actin beta	7p22.1	Hs01060665_g1	2-3	NM_001101
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	12p13.31	Hs02786624_g1	7	NM_001256799
KCNJ6	Potassium voltage-gated channel subfamily J member 6	21q22.13	Hs01040524_m1	3-4	NM_002240
SYP	Synaptophysin	Xp11.23	Hs00300531_m1	3-4	NM_003179
CAMK2A	Calcium/calmodulin-dependent protein kinase II	5q32	Hs00947041_m1	17-18	NM_015981
PAX6	Paired box 6	11p13	Hs01088114_m1	7-8	NM_000280
GRIN1	Glutamate ionotropic receptor NMDA type subunit 1	9q34.3	Hs00609557_m1	1-2	NM_000832

Table S3: Contents and file structure of Python scripts. The file structure will be maintained if the Auto-qPCR program is downloaded from GitHub and run locally. These files will be found inside the 'website' folder if the GitHub repo is pulled or the zip file is downloaded. Folder Name indicates the parent folder and the subfolder containing the program files. File name indicates the file name for each Python script and Function indicates what processes are performed by each script.

Folder Name	File name	Function
Auto-qPCR	main.py	calls app
application	AUTOqPCR.py	inputs data
		inputs conditions
		removes outliers
		calls model
	absolute.py	runs normalization for absolute model
	relative.py	runs relative quantification with delta-CT normalization
	stability.py	runs relative quantification with delta-delta-CT normalization and genomic instability test
	plot.py	creates all graphs
	statistics.py	runs all statistics
	regex_rename.py	function to allow flexible naming
application/template	all html interface files	creates the web form

Table S4: List of all the user inputs for the Auto-qPCR program and purpose of the expected user inputs. Section indicates the spot in the web app where the input box is located. User Input indicates the input box or options as they appear in the web app. Selections and Values indicates possible options for the user to select and the purpose of the input.

Section	User Input	Selections and Values
Main	Select model	Choose the analysis model to run
	Upload your data	Select your csv files
	File information	Choose yes if your file doesn't contain gene names or you want to filter out data from a second probe.
Options	Endogenous control	Genes/targets for normalization
	Cut-off	The threshold for which the standard deviation is above and outliers from technical replicates will be removed. Default = 0.3
	Max Outliers	The proportion of replicates that can be removed. Default = 0.5. With 0.5, if there are 3 replicates, only 1 can be removed
	Preserve highly variable replicates	If set to yes, a second condition is added before a replicate is removed. The difference between the mean and median must be greater than 10 % of the mean
	Calibrator/reference sample	This is the gene/target that is the second normalization in the ΔΔCT model
Visualization Options	Target order	Genes are entered in the order they will appear on the graph
	Sample order	Sample names are entered in the order they will appear on the graph
	Columns for statistics	If a group column is present in the raw data, it must be indicated here to be available for the statistics

Table S5: Description of the statistical tests using each possible selection criteria. The number of groups to compare, '#G' indicates the number of conditions to compare with the variables. The number of variables, '#Var' indicates the number of experimental conditions to compare. The distribution of the data determines if a parametric test will be used, for normally distributed data, or a non-parametric test will be used by the software. 'Measure' indicates if the data was collected on independent samples or on the same samples at different time points. 'Test' indicates the name of the test used by the software based on the user's sections from the other four criteria. Auto-qPCR always uses the same post-hoc test except when only two groups are being compared and no post-hoc test is performed.

# G	# V	Distribution	Measure	Test	Posthoc
2	1	parametric (normal)	Independent	student t-test two tailed, un-paired	none
2	1	parametric (normal)	Repeated measures (dependent)	student t-test two tailed, paired	none
2	1	non-parametric	Independent	Wilcoxon test	none
2	1	non-parametric	Repeated measures (dependent)	Mann-Whitney U test	none
> 2	1	parametric (normal)	Independent	one-way ANOVA	pairwise t-tests with FDR correction
> 2	1	parametric (normal)	Repeated measures (dependent)	one-way ANOVA	pairwise t-tests with FDR correction
> 2	1	non-parametric	Independent	Kruskal-Wallis test	pairwise t-tests with FDR correction
> 2	1	non-parametric	Repeated measures (dependent)	Friedman test	pairwise t-tests with FDR correction
> 2	2	parametric (normal)	Independent	two-way ANOVA	pairwise t-tests with FDR correction
> 2	2	parametric (normal)	Repeated measures (dependent)	two-way ANOVA	pairwise t-tests with FDR correction, for conditions 1,2 and the interaction

Table S6: Results of Auto-qPCR summary output found in summary_data.csv. The DNA region is indicated in Target Name, cell lines are indicated in Sample Name, Indel indicates if there is a duplication or deletion event calculated by the web app, Rep is the number of technical replicates included for analysis, RQ is the relative quantification, Std is the standard deviation and SEM is the standard error of the mean. RQ values from the technical replicates.

Target Name	Sample Name	Indel	Rep	RQ	Std	SEM
CHR1	GM25953	Normal	3	0.958	0.028	0.016
CHR1	GM25975	Normal	3	1.009	0.036	0.021
CHR1	GM25974	Normal	3	1.026	0.011	0.006
CHR1	GM25952	Normal	3	0.962	0.058	0.033
CHR1	Normal	Normal	3	1.000	0.032	0.019
CHR4	GM25953	Normal	3	1.000	0.006	0.003
CHR4	GM25975	Normal	3	1.000	0.012	0.007
CHR4	GM25974	Normal	3	1.000	0.016	0.009
CHR4	GM25952	Normal	3	1.000	0.024	0.014
CHR4	Normal	Normal	3	1.000	0.017	0.010
CHR8	GM25953	Normal	3	1.026	0.035	0.020
CHR8	GM25975	Normal	3	1.027	0.053	0.031
CHR8	GM25974	Normal	3	1.102	0.006	0.003
CHR8	GM25952	Normal	3	1.007	0.028	0.016
CHR8	Normal	Normal	3	1.000	0.009	0.005
CHR10	GM25953	Normal	3	0.913	0.040	0.023
CHR10	GM25975	Normal	3	0.998	0.024	0.014
CHR10	GM25974	Normal	3	0.976	0.044	0.026
CHR10	GM25952	Normal	3	0.979	0.061	0.035
CHR10	Normal	Normal	3	1.000	0.008	0.005
CHR12	GM25953	Normal	3	0.935	0.038	0.022
CHR12	GM25975	Normal	3	1.094	0.005	0.003
CHR12	GM25974	Normal	3	1.140	0.023	0.013
CHR12	GM25952	Normal	3	1.080	0.053	0.031
CHR12	Normal	Normal	3	1.000	0.012	0.007
CHR17	GM25953	Normal	3	0.921	0.054	0.031
CHR17	GM25975	Normal	3	1.061	0.061	0.035
CHR17	GM25974	Normal	3	1.220	0.041	0.024
CHR17	GM25952	Normal	3	1.088	0.202	0.116
CHR17	Normal	Normal	3	1.001	0.049	0.028
CHR18	GM25953	Normal	3	0.938	0.021	0.012
CHR18	GM25975	Normal	3	0.991	0.028	0.016
CHR18	GM25974	Normal	3	0.972	0.032	0.019
CHR18	GM25952	Normal	3	0.988	0.010	0.006
CHR18	Normal	Normal	3	1.000	0.015	0.009
CHR20	GM25953	Normal	3	0.992	0.045	0.026
CHR20	GM25975	Normal	3	1.104	0.007	0.004
CHR20	GM25974	Normal	3	0.927	0.025	0.014
CHR20	GM25952	Normal	3	0.874	0.021	0.012
CHR20	Normal	Normal	3	1.000	0.037	0.021
CHRX	GM25953	Normal	3	0.963	0.030	0.018
CHRX	GM25975	Normal	3	0.931	0.019	0.011
CHRX	GM25974	Normal	3	0.975	0.007	0.004
CHRX	GM25952	Normal	3	0.985	0.069	0.040
CHRX	Normal	Normal	3	1.000	0.027	0.016

Table S7: Statistical results for the absolute quantification found in file ANOVA_results.csv. Target Name indicates the genes compared, DF: degrees of freedom, F is the statistic to determine the p-value, MS: mean squares, SS: sums of squares, measure indicates if the tests were dependent measures for example, in a time course, where cell lines were matched across samples. Dist indicates the distribution is normal (parametric).

Target Name	DF	F	MS	SS	p-value	p-value corrected	Measure	Dist
GAPDH	3	5.491	0.046	0.137	0.00951	0.04753	dependent	parametric
ACTB	3	6.958	0.038	0.115	0.00372	0.01859	dependent	parametric
KCNJ6	3	22.923	20.729	62.188	0.00001	0.00004	dependent	parametric
SYP	3	114.917	58.478	175.433	0.00000	0.00000	dependent	parametric
GRIA1	3	11.24	10.081	30.243	0.0004	0.00201	dependent	parametric

Table S8: Post-hoc results from the statistical analysis of the absolute quantification from the one-way ANOVA. These results are found in file Posthoc_result.csv. The comparisons between individual stages for each gene is show. Target Name indicates the gene of interest. **A and B** show the two groups being compared. DF: degrees of freedom, p-value correct is the value corrected for multiple comparisons, p-value before correction for a paired t-test. Parametric, True means a normal distribution was selected.

Target Name	A	B	DF	p-value corrected	p-value	Paired	Parametric
KCNJ6	IPSC	NPC	5	0.85667	0.73431	TRUE	TRUE
KCNJ6	IPSC	DA4W	5	0.00845	0.00282	TRUE	TRUE
KCNJ6	IPSC	DA6W	5	0.00845	0.00253	TRUE	TRUE
KCNJ6	NPC	DA4W	5	0.01157	0.00705	TRUE	TRUE
KCNJ6	NPC	DA6W	5	0.01157	0.00771	TRUE	TRUE
KCNJ6	DA4W	DA6W	5	0.85667	0.85667	TRUE	TRUE
SYP	IPSC	NPC	5	0.18543	0.171	TRUE	TRUE
SYP	IPSC	DA4W	5	0.0001	0.00002	TRUE	TRUE
SYP	IPSC	DA6W	5	0.00018	0.00009	TRUE	TRUE
SYP	NPC	DA4W	5	0.00018	0.00012	TRUE	TRUE
SYP	NPC	DA6W	5	0.00018	0.00011	TRUE	TRUE
SYP	DA4W	DA6W	5	0.18543	0.18543	TRUE	TRUE
GRIA1	IPSC	NPC	5	0.06779	0.05649	TRUE	TRUE
GRIA1	IPSC	DA4W	5	0.03575	0.01192	TRUE	TRUE
GRIA1	IPSC	DA6W	5	0.03575	0.01137	TRUE	TRUE
GRIA1	NPC	DA4W	5	0.06779	0.03449	TRUE	TRUE
GRIA1	NPC	DA6W	5	0.06779	0.0519	TRUE	TRUE
GRIA1	DA4W	DA6W	5	0.35174	0.35174	TRUE	TRUE

Table S9: Example of output from the relative delta-CT analysis from the file clean_data.csv showing the top 10 rows of data. Target Name indicates the gene analyzed, Sample Name indicates the cell line, rq is the relative quantification for each replicate, rq-mean is the mean value of the replicates, rqSD is the standard deviation of the replicates, rqSEM is the standard error of the replicates, Outliers indicates if each outlier is a replicate, Group indicates the group used for statistics for the summary data.

Target Name	Sample Name	rq	rqMean	rqSD	rqSEM	Outliers	Group
PAX6	AIW002-2-	0.0187	0.0223	0.0032	0.0018	FALSE	D0
PAX6	AIW002-2-	0.0248	0.0223	0.0032	0.0018	FALSE	D0
PAX6	AIW002-2-	0.0235	0.0223	0.0032	0.0018	FALSE	D0
PAX6	AIW002-2-	0.0072	0.0073	0.0004	0.0002	FALSE	D7
PAX6	AIW002-2-	0.0069	0.0073	0.0004	0.0002	FALSE	D7
PAX6	AIW002-2-	0.0077	0.0073	0.0004	0.0002	FALSE	D7
PAX6	KYOU--	0.1261	0.1193	0.0065	0.0038	FALSE	D0
PAX6	KYOU--	0.1131	0.1193	0.0065	0.0038	FALSE	D0
PAX6	KYOU--	0.1187	0.1193	0.0065	0.0038	FALSE	D0
PAX6	KYOU--	0.0202	0.0210	0.0007	0.0004	FALSE	D7

Table S10: Statistical results from the relative quantification comparing the delta-CT and delta-delta-CT using student t-tests. Target name indicate the gene being compared, DF: degrees of freedom, tail; two tail t-test, paired FALSE indicated an unpaired t-test. The p-values are shown under p-val. Model indicates if the delta-CT or the delta-delta-CT method was used.

Target Name	DF	T	tail	paired	p-value	model	effect size	power	Bayes factor
PAX6	1	1.361	two-sided	FALSE	0.40342	delta CT	1.449	0.129	0.847
CAMK2A	1	-3.277	two-sided	FALSE	0.18855	delta CT	1.405	0.125	1.359
GRIN1	1	-3.744	two-sided	FALSE	0.16616	delta CT	1.836	0.162	1.454
PAX6	1	1.361	two-sided	FALSE	0.40342	delta delta CT	1.449	0.129	0.847
CAMK2A	1	-3.277	two-sided	FALSE	0.18855	delta delta CT	1.405	0.125	1.359
GRIN1	1	-3.744	two-sided	FALSE	0.16616	delta delta CT	1.836	0.162	1.454

Table S11: Manual processing compared to Auto-qPCR processing with a range of cut-off values for std to exclude replicates, with or without preserving highly variable outliers. Calculations are all using the absolute model to quantify NRXN3 expression with and without cocaine treatment in three brain regions. Values that differ across processing conditions are highlighted in bold. **Left**, the sample information for Region, Treatment and code name of each mouse (biological replicate) are listed. The processing methods, Manual or Auto-qPCR, are labelled. The std cut-off is the value for which std exceeded for outliers to be moved. The settings for preserving highly variable technical if the ration of mean-media/media is less than 0.1 is indicated by 'yes'. RNA indicates the RNA quantification values.

			Manual	Auto-qPCR				
Preserve high variation replicates			yes	no	no	no	yes	yes
Std cut-off			0.29	0.3	0.275	0.2	0.3	0.275
Region	**Treatment**	**Mouse**	**RNA**	**RNA**	**RNA**	**RNA**	**RNA**	**RNA**
STN	Saline	B4bis	0.2564	0.2564	0.2564	0.2817	0.2564	0.2564
STN	Saline	B6	0.1933	0.1933	0.1933	0.1933	0.1933	0.1933
STN	Saline	R6	0.3290	0.3290	0.3290	0.3055	0.3290	0.3290
STN	Saline	V3	0.2845	0.2845	0.2845	0.3357	0.2845	0.2845
STN	Saline	V4	0.3259	0.3259	0.3259	0.3259	0.3259	0.3259
STN	Cocaine	R5Bis	0.4570	0.4570	0.4570	0.4570	0.4116	0.4116
STN	Cocaine	R6bis	0.1708	0.1708	0.1708	0.1708	0.1708	0.1708
STN	Cocaine	R8bis	0.4253	0.4253	0.4253	0.4253	0.4253	0.4253
STN	Cocaine	V2	0.2538	0.1659	0.1659	0.1659	0.1987	0.1987
STN	Cocaine	V8	0.1818	0.1818	0.1818	0.1818	0.1818	0.1818
GP	Saline	B4bis	0.2541	0.2541	0.2541	0.2541	0.2541	0.2541
GP	Saline	R6	0.7107	0.7107	0.7107	0.7107	0.7107	0.7107
GP	Saline	V3	0.4125	0.4125	0.4125	0.4125	0.4125	0.4125
GP	Saline	V4	0.2991	0.2991	0.2991	0.2991	0.2991	0.2991
GP	Cocaine	R5Bis	0.5021	0.5021	0.5021	0.5021	0.4988	0.4988
GP	Cocaine	R6bis	0.9500	0.9500	0.9500	0.9500	0.9500	0.9500
GP	Cocaine	R8bis	1.0169	1.0169	1.0169	0.9455	1.0169	1.0169
GP	Cocaine	V2	0.8538	0.8538	0.8538	0.7797	0.8538	0.8538
GP	Cocaine	V8	0.9486	0.9486	0.9486	0.9486	0.9486	0.9486
SN	Saline	B4bis	0.7854	0.8745	0.7854	0.7854	0.7317	0.7317
SN	Saline	R6	0.8751	1.0784	1.0784	1.0784	0.8751	0.8751
SN	Saline	V3	1.3306	1.3306	1.3306	1.3306	1.1553	1.1553
SN	Saline	V4	1.0575	1.0575	1.0575	1.0575	0.8940	0.8940
SN	Cocaine	R5Bis	1.2379	1.2379	1.2379	1.2379	1.2379	1.2379
SN	Cocaine	R6bis	1.0016	1.1607	1.1607	1.2982	1.0016	1.0016
SN	Cocaine	R8bis	0.4393	0.4393	0.4393	0.4393	0.4393	0.4393
SN	Cocaine	V2	1.0196	1.0196	1.0196	1.0971	1.0196	1.0196
SN	Cocaine	V8	2.2979	2.2979	2.2979	2.2979	2.2979	2.2979

Table S12: Comparison of variance between manual processing and Auto-qPCR. The variance between RNA quantity values calculated manually or with Auto-qPCR were calculated between each mean value found in table S11. For each brain region the sum of the variance was calculated. The same comparison was performed between manual processing and Auto-qPCR with the standard cut-off of 0.3 and the standard cut-off together plus the preserve extreme values option.

Region	Cut-off 0.3	Cut-off 0.3 + Preserve
STN	0.004	0.003
GP	0.000	0.000
SN	0.037	0.030
All regions	0.037	0.033

Table S13: Results of the statistical analysis of control vs. cocaine treatment for all brain regions. The results of an unpaired, two tailed students t-test performed in Auto-qPCR using the statistic selections. The table is found in the file 'ttest_results.csv'.

Target Name	DF	T	tail	paired	p-val
B2M	18.54	0	two-sided	FALSE	1.000
NRXN3	22.74	-1.555	two-sided	FALSE	0.134

Table S14: Posthoc results after one-way ANOVA comparing brain regions. Control and cocaine treatment samples were pooled together. Target name indicates the gene tested. **A and B** indicate the two regions being compared. DF: degree of freedom.

Target Name	A	B	DF	p-value corrected	p-value	correction method
B2M	STN	GP	9	1.0000	1.0000	fdr_bh
B2M	STN	SN	12	1.0000	1.0000	fdr_bh
B2M	GP	SN	16	1.0000	1.0000	fdr_bh
NRXN3	STN	GP	9	0.0071	0.0047	fdr_bh
NRXN3	STN	SN	8	0.0048	0.0016	fdr_bh
NRXN3	GP	SN	16	0.0549	0.0549	fdr_bh

Table S15: One-way ANOVA and posthoc test comparing groups of brain region and treatment. The ANOVA results are shown for both B2M and NRXN3 for the overall effect of treatment and brain region together (One-way ANOVA. The post-hoc tests for the relevant comparisons are shown for each brain region with and without cocaine treatment (post-hoc).

Target Name	Comparison	DF	p-value corrected	p-value	Test
B2M	Treatment and region	5	0.3885	0.7771	One-way ANOVA
NRXN3	Treatment and region	5	0.0006	0.0011	One-way ANOVA
NRXN3	STN_Control vs STN_Cocaine	8	0.9977	0.9977	post-hoc
NRXN3	GP_Control vs GP_Cocaine	7	0.0176	0.0334	post-hoc
NRXN3	SN_Control vs SN_Cocaine	5	0.4127	0.4762	post-hoc

Table S16: Two-way ANOVA and posthoc tests comparing brain region, treatment and interaction. The relevant information was selected from the output files 'ANOVA_results.csv' and 'Posthoc_results.csv'. The 2-way ANOVA results are shown for NRXN3 for the overall effect of brain region (Group1), treatment (Group2) and the interaction effect of region and treatment (Group1*Group2) (upper table). The post-hoc tests for the relevant comparisons are shown for each brain region with and without cocaine treatment for each brain region indicated under contrast. The 2-way ANOVA results are shown on top and the post-hoc multiple t-test comparisons are shown on the bottom, indicated in the Test column

Target Name	Contrast	DF	p-value corrected	p-value	Test
NRXN3	Group1: Region	2	0.0004	0.0001	ANOVA
NRXN3	Group2: Treatment	1	0.2265	0.0755	ANOVA
NRXN3	Group1 * Group2	2	1.0000	0.3513	ANOVA
NRXN3	all: Control vs Cocaine	23	NA	0.1337	post-hoc
NRXN3	STN: Control vs Cocaine	8	0.9977	0.9977	post-hoc
NRXN3	GP: Control vs Cocaine	7	0.0529	0.0176	post-hoc
NRXN3	SN: Control vs Cocaine	5	0.6190	0.4127	post-hoc