

Adaptation pédagogique d'un pipeline RNAseq par Devin

journée « IA & Scripting »
13 juin 2025

Contexte

- Module du M2 de génétique sur les analyses de RNAseq bulk
 - Depuis des années sur un (très) petit dataset pour pouvoir tourner sans crasher
 - Mais maintenant gros serveur dédié à l'enseignement * :



Plateforme d'e-Learning pour l'Analyse de données Scientifiques Massives

- => adaptation des analyse en jupyter notebooks
Mais toujours sur le petit dataset....



* Avec Claire Vandiedonck et Pierre Poulain

Contexte

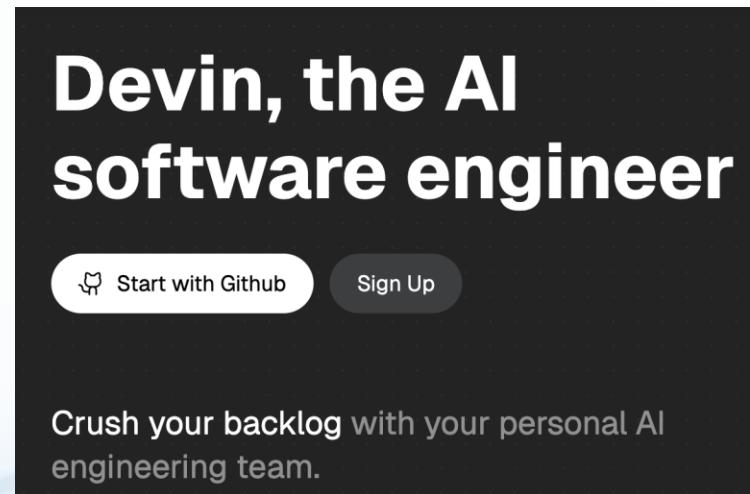
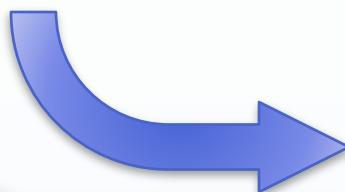
- Formation pour le CEA en 2023 et 2024, montée sur le cluster de l'IFB *
 - Toujours en jupyter notebooks, TRES explicites et adaptables
 - Passage sur nouveau dataset



* Avec Claire Vandiedonck

Adaptation des notebooks

- Formation CEA :
 - Jupyter notebooks très « verbeux »
 - Beaucoup de variables pour être réutilisables
- Simplification pour Plasma :
 - Jupyter notebooks avec moins de détails
 - Strict minimum de variables
 - Tout en anglais
- Mais 11 notebooks !!



The image shows the landing page for Devin, the AI software engineer. The background is dark with white text. At the top, it says "Devin, the AI software engineer". Below that are two buttons: "Start with Github" and "Sign Up". At the bottom, it says "Crush your backlog with your personal AI engineering team."

Interface Devin

Update RNAseq Course Notebooks

Hi Devin

In addition to my research work, I'm an associated professor in genomics and bioinformatics.

I have to update a one-week course on RNAseq bioinformatic analysis.

I want to adapt the jupyter notebooks that are available on the github repo of my colleague Claire, here: https://github.com/CVandiedonck/CEA_rnaseq.git

for students in their second year of studies, so all the

Chat GPT-like

The analysis pipeline is in two parts:

- the first part deals with steps in bash (from Pipe00b notebook to Pipe_06 notebook)
- the second part deals with steps in R (from Pipe_08 notebook to Pipe_11 notebook)

I want to simplify these notebooks, in particular by removing most of the variables in the code.

The data to be analysed are going to be stored on the Plasmabio serveur, in the /srv/data/meg-m2-rnaseq/genome/ and in /srv/data/meg-m2-rnaseq/experimental_data folders.

Students will work by connecting to a Jupyter Hub with a login and a password, and will be in a controlled environnement launched by Jupyter Hub.

Currently, we have two different environments to run the book and the

Follow Devin

Timelapse

IDE

Progress Shell Browser Editor Planner

001 check_for_repo("RNAseq_Plasma")
002 clone_repo("https://github.com/CVandiedonck/CEA_rnaseq.git")
003 read_readme_and_gather_information()
004 find_and_create_todo_list()
005 review_and_simplify_text_and_code()
006 write_simplified_notebooks()
> 007 report_to_user()

Current Step: 007 report_to_user()

Partie « système »

Step	Outcome	Time Taken	Timestamp
> 001 create.github_re...	●	11:05	Oct 30, 2024, 17:29
> 002 clone_repo("http...	●	01:54	Oct 30, 2024, 17:31
> 003 read_readme_and...	●	07:39	Oct 30, 2024, 17:39
> 004 find_and_create_...	●	00:15	Oct 30, 2024, 17:39
> 005 review_text_and_...	●	00:40	Oct 30, 2024, 17:40
> 007 summarize_notebo...	●	00:21	Oct 30, 2024, 17:40
> 008 define_minimum_v...	●	00:26	Oct 30, 2024, 17:40
> 009 suggest_simplifi...	●	00:11	Oct 30, 2024, 17:41
> 010 mark_in_todo_lis...	●	00:15	Oct 30, 2024, 17:41
> 006 review_text_and_...	●	00:43	Oct 30, 2024, 17:41
> 007 summarize_notebo...	●	00:25	Oct 30, 2024, 17:42
> 008 define_minimum_v...	●	00:40	Oct 30, 2024, 17:43

Prompt initial : (oui, je suis bavarde 😊)

Contexte :

I have to update a one-week course on RNAseq bioinformatic analysis.

I want to adapt the jupyter notebooks that are available on the github repo of my colleague Claire, here: https://github.com/CVandiedonck/CEA_rnaseq.git

for students in their second-year in a Master of genetics.

Précisions et demande explicite :

The master is international, so we have foreign students, so all the teaching material has to be in english.

The analysis pipeline is in two parts:

- the first part deals with steps in bash (from Pipe00b notebook to Pipe_06 notebook)
- the second part deals with steps in R (from Pipe_08 notebook to Pipe_11 notebook)

I want to simplify these notebooks, in particular by removing most of the variables in the code.

Prompt initial :

Infos sur le systeme de destination

The data to be analysed are going to be stored on the Plasmabio serveur, in the /srv/data/meg-m2-rnaseq/genome/ and in /srv/data/meg-m2-rnaseq/experimental_data folders.

Students will work by connecting to a Jupyter Hub with a login and a password, and will be in a controlled environnement launched by Jupyter Hub.

Currently we have two different environments, one for the bash part, and the second one for the R part.

The working directory will have the form:

/srv/home/scaburet/meg_m2_rnaseq_bash

or

/srv/home/scaburet/meg_m2_rnaseq_r

The environments are built on the Plasmabio server using a yml file and a postBuild file to install the various programs and libraries.

Prompt initial :

Détails des tâches demandées

Could you please:

- create a repo on my github account named RNAseq_Plasma
- clone the jupyter notebooks present in Claire's CEA_rnaseq repo into this new one, but taking only the ones from this year (except the Pipe_00a-bash-bienvenue-sur-IFB.ipynb one, because we won't be in the same serveur, nor the Pipe_01-bash_download-files.ipynb because the data will be already present on the server)
- review the text and the code in all the notebooks
- summarize briefly what each notebook does
- define the minimum number of variables that we would need in the simplified version of the pipeline
- suggest the various simplifications that you would see fit (Claire and I will decide which ones of your suggested simplifications will be applied)
- write an example of a simplified version of the Pipe_02-bash_raw-data-quality.ipynb

Do you need any further info to start this task ?

Partie « systeme » : Progress

The screenshot shows a software interface with a dark theme. At the top, there is a navigation bar with several items: "Follow Devin" (with a follow icon), "Timelapse" (with a timelapse icon, circled in red), "IDE" (with a code editor icon), and a three-dot menu icon. Below the navigation bar, there is a horizontal menu with icons: "Progress" (circled in red), "Shell", "Browser", "Editor", and "Planner". The main area of the interface displays a code editor with the following Python-like pseudocode:

```
> 001  create_github_repo("RNAseq_Plasma")
002  clone_repo("https://github.com/CVandiedonck/CEA_rnaseq.git")
003  read_readme_and_gather_information()
004  find_and_create_todo_from_notebooks()
005  if notebook = next(todo_list):
006      review_text_and_code(notebook)
007      summarize_notebook(notebook)
008      define_minimum_variables(notebook)
009      suggest_simplifications(notebook)
010      mark_in_todo_list(notebook)
011      goto 005
012  write_simplified_example("Pipe_02-bash_raw-data-quality.ipynb")
013  report_to_user()
```

Gestion des identifiants et accès

❖ I understand you'd like me to help adapt Claire's RNAseq course materials for your international Master's students, including creating a new GitHub repo, reviewing and simplifying the notebooks, and providing recommendations for improvements. **Before I begin, I'll need access to create the RNAseq_Plasma repository on your GitHub account.**

sc My github username is [Scaburet](#)

How do I provide you with the personal access token with repository creation permissions ?

❖ I can help guide you through creating and sharing a personal access token securely. Go to GitHub Settings > Developer Settings > Personal Access Tokens > Tokens (classic) and generate a new token with 'repo' permissions. For security, I'll provide a secure form for you to share the token once generated.

🔑 [Provide Secrets](#)

Send secrets

Key	Value	Sensitive?
Token_for_Devin	<input checked="" type="checkbox"/> (-)

(+) (↑) (≡) Send

Partie « systeme » : Progress

⌚ Progress ⏺ Shell ⏻ Browser </> Editor ⏷ Planner

Executed # Create a directory for the ...

```
ubuntu@devin-box:~/workspace/rnaseq$ # Create a directory for the filtered notebooks
ubuntu@devin-box:~/workspace/rnaseq$ mkdir -p selected_notebooks && \
> # Copy only the notebooks we want (excluding Pipe_00a and Pipe_01, and selecting the co
rrect versions)
> cp Pipe_00b-bash_intro-to-bash.ipynb selected_notebooks/ && \
> cp Pipe_02-bash_raw-data-quality.ipynb selected_notebooks/ && \
> cp Pipe_03-bash_preprocessing-and-check.ipynb selected_notebooks/ && \
> cp Pipe_04-bash_classical-reads-mapping.ipynb selected_notebooks/ && \
> cp Pipe_05-bash_mapping-quality.ipynb selected_notebooks/ && \
> cp Pipe_06-bash_reads-counts.ipynb selected_notebooks/ && \
> cp Pipe_08-R403_counts-exploratory-analysis-I.ipynb selected_notebooks/ && \
> cp Pipe_09-R403-DESeq2-normalisation-DE.ipynb selected_notebooks/ && \
> cp Pipe_10-R403-Normcounts-exploratory-analysis-II.ipynb selected_notebooks/ && \
> cp Pipe_11-R403-ORA-GSEA.ipynb selected_notebooks/ && \
> # List the selected notebooks to verify
> ls -la selected_notebooks/
total 18
drwxrwxr-x 2 ubuntu ubuntu      12 Oct 30 16:31 .
drwxrwxr-x 4 ubuntu ubuntu      33 Oct 30 16:31 ..
-rw-rw-r-- 1 ubuntu ubuntu 188590 Oct 30 16:31 Pipe_00b-bash_intro-to-bash.ipynb
-rw-rw-r-- 1 ubuntu ubuntu  32354 Oct 30 16:31 Pipe_02-bash_raw-data-quality.ipynb
-rw-rw-r-- 1 ubuntu ubuntu  45260 Oct 30 16:31 Pipe_03-bash_preprocessing-and-check.ipyn
b
-rw-rw-r-- 1 ubuntu ubuntu  52342 Oct 30 16:31 Pipe_04-bash_classical-reads-mapping.ipyn
b
-rw-rw-r-- 1 ubuntu ubuntu  45278 Oct 30 16:31 Pipe_05-bash_mapping-quality.ipynb
-rw-rw-r-- 1 ubuntu ubuntu  33760 Oct 30 16:31 Pipe_06-bash_reads-counts.ipynb
-rw-rw-r-- 1 ubuntu ubuntu  52376 Oct 30 16:31 Pipe_08-R403_counts-exploratory-analysi
I.ipynb
-rw-rw-r-- 1 ubuntu ubuntu  72438 Oct 30 16:31 Pipe_09-R403-DESeq2-normalisation-DE.ipyn
b
-rw-rw-r-- 1 ubuntu ubuntu  92287 Oct 30 16:31 Pipe_10-R403-Normcounts-exploratory-analy
sis-II.ipynb
-rw-rw-r-- 1 ubuntu ubuntu 1708020 Oct 30 16:31 Pipe_11-R403-ORA-GSEA.ipynb
ubuntu@devin-box:~/workspace/rnaseq$
```

Suivi de toutes les commandes en temps réel ou après coup

Partie « systèmes » : Shell

The screenshot shows a terminal window with several tabs at the top: Progress, Shell (which is selected and highlighted with a red circle), Browser, Editor, and Planner. The main pane displays a command-line session. A large red circle highlights the command `pip install nbconvert && \` and the subsequent steps for converting notebooks. Another red circle highlights the command `pip install nbconvert && \` in the command history table.

```
-rw-rw-r-- 1 ubuntu ubuntu 1700000  Oct 30 10:51 Pipe_01-R403_GA_CSE.ipynb
ubuntu@devin-box:~/workspace/rnaseq$ # Install nbconvert if not already installed
ubuntu@devin-box:~/workspace/rnaseq$ pip install nbconvert && \
> # Convert the intro notebook to script format for examination
> jupyter nbconvert --to markdown selected_notebooks/Pipe_00b-bash_intro-to-bash.ipynb && \
> # Also convert one R notebook to understand the R environment
> jupyter nbconvert --to markdown selected_notebooks/Pipe_08-R403_counts-exploratory-analysis-I.ipynb & \
> # List the converted files
> ls -la selected_notebooks/*.md
Requirement already satisfied: nbconvert in /home/ubuntu/.local/lib/python3.12/site-packages (7.16.4)
)
Requirement already satisfied: beautifulsoup4 in /home/ubuntu/.local/lib/python3.12/site-packages (from nbconvert) (4.12.3)
Requirement already satisfied: bleach!=5.0.0 in /home/ubuntu/.local/lib/python3.12/site-packages (from nbconvert) (6.2.0)
Requirement already satisfied: defusedxml in /home/ubuntu/.local/lib/python3.12/site-packages (from nbconvert) (0.7.1)
Requirement already satisfied: jinja2>=3.0 in /home/ubuntu/.local/lib/python3.12/site-packages (from nbconvert) (3.1.4)
```

Command History	Duration	Start Timestamp
# Create a new directory for claire's repo mkdir -p CEA_rnaseq ... 00:00 Oct 30, 2024, 17:...		
# Clean up the failed clone attempt and create a fresh director... 00:24 Oct 30, 2024, 17:...		
# Try to get repository information using GitHub CLI with our t... 00:00 Oct 30, 2024, 17:...		
# Clone the repository using our token git clone https://oauth2... 00:02 Oct 30, 2024, 17:...		
# Create a directory for the filtered analysis pipeline select ... 00:00 Oct 30, 2024, 17:...		
# Install nbconvert if not already installed pip install nbconv... 00:04 Oct 30, 2024, 17:...		
# Read the contents of both converted notebooks echo --- Bash ... 00:00 Oct 30, 2024, 17:...		
# Create a summary of required packages and environment setup ec... 00:00 Oct 30, 2024, 17:33		
# Convert the notebook to markdown for review jupyter nbconvert ... 00:00 Oct 30, 2024, 17:39		
jupyter nbconvert --to markdown "selected_notebooks/Pipe_02-bash... 00:00 Oct 30, 2024, 17:41		
jupyter nbconvert --to markdown "selected_notebooks/Pipe_03-bash... 00:00 Oct 30, 2024, 17:43		
mkdir -p selected_notebooks cd selected_notebooks gh repo clone ... 00:00 Oct 30, 2024, 17:46		

Partie « systeme » : Editor

The screenshot shows a Jupyter Notebook interface with the following details:

- Top Bar:** Progress, Shell, Browser, Editor (highlighted by a red circle), Planner.
- Left Sidebar (EXPLORER):** Shows a file tree structure:
 - ubuntu
 - todo.txt
 - workspace
 - rnaseq
 - Pipe_00b-bash_intro-to-bash.simplified.md
 - Pipe_00b-summary.md
 - Pipe_00b-variables.md
 - Pipe_00b-suggestions.md
 - Pipe_02-bash_raw-data-quality.md
 - Pipe_02-t
- Code Cell (Main Area):** Displays the content of the notebook `Pipe_02-summary.md`. The content describes a pipeline for quality control analysis on raw RNA sequencing data using FastQC and MultiQC tools. It includes sections on purpose, key functions, technical details, prerequisites, role in pipeline, and learning objectives. A red oval highlights the first few lines of the purpose section and the entire learning objectives section at the bottom.

Partie « systeme » : Suggested changes



Suggested Simplifications for Pipe_03

1. Language and Documentation

Current Issues:

- Mixed French and English text
- Complex technical explanations
- Inconsistent documentation style

Suggestions:

- ✓ 1. Translate all content to English
 - Convert all French text, comments, and explanations
 - Use clear, simple English suitable for international students
 - Include a glossary of key terms
- ✓ 2. Restructure Documentation
 - Add clear learning objectives at start
 - Break down complex concepts into digestible sections
 - Use consistent formatting for commands and explanations
 - Add more visual aids (diagrams, flowcharts)

Suggested changes



2. Variable Structure

3. Code Organization

4. Quality Control Process

5. Educational Improvements

6. Technical Optimizations

7. Specific Code Changes

8. Implementation Priority

1. High Priority

- English translation
- Variable simplification
- Directory structure standardization

2. Medium Priority

- Code modularization
- QC parameter standardization
- Documentation improvements

3. Low Priority

- Additional visualizations
- Optional exercises
- Advanced optimizations

Benefits

1. For Students

- Clearer learning path
- Better understanding of concepts
- More hands-on practice

2. For Instructors

- Easier maintenance
- Consistent structure
- Better assessment opportunities

3. For System

- Reduced complexity
- Better resource usage
- Easier troubleshooting

Résultat de la première session :



I have completed all requested tasks for adapting the RNA-seq course materials. Here's a summary of the deliverables:

1. Pipeline Simplification Report (pipeline_simplification_report.md)

- Summaries of all notebooks (Pipe_00b to Pipe_06 and Pipe_08 to Pipe_11)
- Minimum required variables for both bash and R parts
- Suggested simplifications for each notebook
- Implementation recommendations
- Next steps for deployment

- ▼ Pipe_02-bash_raw-data-quality.md
- ▼ Pipe_02-summary.md
- ▼ Pipe_02-variables.md
- ▼ Pipe_02-simplifications.md

2. Simplified Example Notebook (Pipe_02-bash_raw-data-quality.simplified.ipynb)

- Minimized variables
- Clear English instructions
- Educational content for Master's students
- Self-assessment questions
- Proper Plasmabio paths

Key Features:

- All content is in English
- Paths adapted for Plasmabio server
- Simplified for second-year Master's students
- Clear separation between bash and R components
- Educational focus with step-by-step instructions

**Création du repo : 17h19
Report to user : 18h15...**

Conclusions

- TRES rapide et TRES efficace
- Intuitif, facile d'utilisation, avec explications, feedback...
- Snapshot pour terminer la session et repartir dessus la fois suivante
- 2ème session : merging et simplification des jupyter en bash, ajustement du niveau de verbosité, scan des jupyter R pour la liste des librairies, écriture des fichiers de config pour les 2 environnements
 - > Devin a parfois un peu perdu le lien avec les infos de la première session
- Utilisé aussi pour un projet de recherche plus large, avec automatisation d'annotation, gestion API, création de bdd et de site web interactif en perspective
- Evidemment payant : 50 \$ / mois pour 40 ACUs
 - > cette session a utilisé 14.6 ACU

Création du snapshot de fin de session :

How can I come back to this session nextime, so we can start from this point?

SC Thank you for the kind words! To continue from this point later, you can create a machine snapshot. Click the menu icon (three dots) in the top right corner of Devin's workspace and select "Create machine snapshot". Give it a name like "RNAseq_Course_Initial" and a description. Next time, you can start a new session by selecting this snapshot from the Devin home page under "Select a snapshot".

SC All right, can you provide me with a description of this session to put in the Snapshot, one that will be useful for you next time?

 Copy  Translate Oct 30, 2024, 18:21

SC Here's a description for the snapshot:

"Initial setup for RNAseq course adaptation. Completed: (1) Review of all notebooks from Pipe_00b to Pipe_11, (2) Created pipeline simplification report with minimum variables and suggested improvements, (3) Generated simplified example of Pipe_02 with educational content. Ready for: Implementation of suggested simplifications and creation of remaining simplified notebooks. Environment includes necessary tools for RNA-seq analysis and Jupyter notebook development."

Infos sur la session

This feature provides an AI-powered analysis of your Devin session, helping you understand what happened, what could be improved, and how to get better results with Devin.

- **Understand Devin:** See a timeline and details of key events, especially where things didn't go smoothly.
- **Actionable Feedback:** Get specific suggestions and configuration changes to help you improve future sessions.
- **Improved Prompt:** Try the suggested prompt to see the effects of frontloading task requirements and pointers.

Beta: We're eager for your feedback—let us know whether these insights help you understand and use Devin better!

Session Insights ⓘ

ACU Usage

14.6

User messages

11

Session size

⚠️ L ⓘ

Analysis

Last updated 5:54 PM

Issue Timeline

Actionable Feedback

Session size guide

Size	ACUs	User messages	Health
XS	≤ 2	and ≤ 2	Healthy
S	> 2	or > 2	Healthy
M	> 5	or > 5	Healthy
L	> 10	or > 10	Unhealthy
XL	> 20	or ≥ 20	Very unhealthy

Tip: to keep sessions small and efficient, provide all important information upfront in the initial prompt

Infos sur la session

Issue Timeline Actionable Feedback X

ISSUES DETECTED

Communication issue (medium)

GitHub access setup required unnecessary back-and-forth communication. Devin asked for access but didn't proactively explain the personal access token process, leading to multiple messages where the user had to ask how to share credentials and Devin had to guide the token creation process. This could have been streamlined by explaining the full access procedure upfront.

Communication issue (low)

Repository access confusion where Devin claimed it couldn't access Claire's repository but was apparently working on the notebooks successfully, creating unnecessary questions from the user about how the work was being done. This was more of a communication clarity issue than a major blocker.

TIMELINE

User provides comprehensive task requirements

User requests GitHub repo creation, notebook adaptation, and comprehensive analysis of RNA-seq course materials (4:18 PM)

Devin requests access without explaining token process

Devin asks for GitHub access but doesn't proactively explain how to create and share personal access token, leading to follow-up questions (4:18 PM)

Token sharing back-and-forth resolved

User successfully creates and shares GitHub personal access token after guidance (4:25 PM)

Repository access confusion emerges

Devin claims inability to access Claire's repository while apparently working on the notebooks, creating user confusion (4:29-4:52 PM)

User clarifies repository boundaries

User reminds Devin to only make changes in new repository, not Claire's original repo (4:51 PM)

Comprehensive deliverables completed

Devin delivers pipeline simplification report and simplified example notebook meeting all requirements (5:14-5:15 PM)

User expresses high satisfaction

User responds enthusiastically: "What you did is impressive! It's great!" confirming successful task completion (5:19 PM)

Session prepared for continuation

User creates snapshot to continue work in future sessions, indicating ongoing value and project success (5:22 PM)

● High impact issue

● Significant event

● Medium impact issue

● Value provided

Infos sur la session > Feedback

Issue Timeline Actionable Feedback

IMPROVED PROMPT

[Start New Session →](#)

Hi Devin

In addition to my research work, I'm an associated professor in genomics and bioinformatics.

I have to update a one-week course on RNAseq bioinformatic analysis.

My GitHub username is Scaburet. I'll provide you with a personal access token with repo permissions after you start - please guide me through creating one when you're ready.

I want to adapt the jupyter notebooks that are available on the github repo of my colleague Claire, here: [https://github.com/CVandiedonck/CEA_rnaseq.git] (https://github.com/CVandiedonck/CEA_rnaseq.git) - please only read from this repo, don't make any changes to Claire's original

CHANGES MADE

- Provide GitHub username upfront and set expectations about token setup process
- Clarify that Claire's repository should be read-only to avoid confusion

No recommended configuration changes

Mémoire d'infos utiles à plusieurs projets

Knowledge Suggestions All Knowledge 

Close

Knowledge are pieces of information that Devin will remember and retrieve in future sessions. Devin automatically suggests new knowledge items or updates to existing knowledge items based on your messages.

Status	Type	Trigger	Folder
New	Use when:	When tasked with adapting educational materials, such as Jupyter notebooks for a course	 Root Folder
<p>To simplify Jupyter notebooks for educational purposes, it is often necessary to remove most of the variables in the code to streamline the learning process and focus on the core concepts being taught. Reviewing the text and code in Jupyter notebooks is essential to ensure that the material is appropriate for the target audience's skill level and that the language used is suitable for all students, including those from international backgrounds.</p>			
Macro: e.g. test-local-webapp			
New	Use when:	When creating a new GitHub repository for a user and requiring a personal access token	 Root Folder
New	Use when:	When working with multiple GitHub repositories in a project	 Root Folder

Détails: Suggested changes

2. Variable Structure

Current Issues:

- Too many variables
- Complex path structures
- Redundant configurations

Suggestions:

1. Simplify Variable Structure

- Reduce to 5 core variables (as defined in variables.md)
- Use consistent naming convention
- Remove redundant path variables

2. Standardize Paths

- Use fixed directory structure
- Simplify path construction
- Remove IFB-specific paths

Détails: : Suggested changes

3. Code Organization

Current Issues:

- Long, complex code blocks
- Redundant operations
- Unclear workflow progression

Suggestions:

1. Modularize Code

- Break long operations into smaller, focused cells
- Add clear cell titles and purposes
- Group related operations

2. Streamline Workflow

- Remove redundant file checks
- Combine similar operations
- Add progress indicators

Détails: : Suggested changes

4. Quality Control Process

Current Issues:

- Complex parameter settings
- Multiple separate QC steps
- Redundant report generation

Suggestions:

1. Simplify QC Parameters

- Use standardized quality thresholds
- Remove unnecessary options
- Add explanations for parameter choices

2. Integrate QC Steps

- Combine fastp and FastQ Screen reports
- Streamline MultiQC generation
- Add quality score interpretation guide

Détails: : Suggested changes

5. Educational Improvements

Current Issues:

- Focus on technical details over concepts
- Limited explanation of biological significance
- Complex tool usage

Suggestions:

1. Add Learning Elements

- Include mini-exercises
- Add interpretation guidelines
- Provide real-world examples

2. Improve Visualization

- Add quality score interpretation examples
- Include sample report analysis
- Create decision-making flowchart

Détails: : Suggested changes

6. Technical Optimizations

Current Issues:

- Manual file management
- Complex logging
- Redundant disk operations

Suggestions:

1. Automate File Management

- Create standard directory structure
- Implement automatic file naming
- Add cleanup operations

2. Simplify Logging

- Standardize log format
- Reduce redundant logging
- Add meaningful progress indicators

Détails: : Suggested changes

```
## 7. Specific Code Changes
### Example Simplifications:
```bash
Original complex code
time for read1 in $(ls "${rawfolder}"*_1.fastq.gz); do
 id=$(basename ${read1} | cut -d"_" -f1)
 read2=$(echo ${read1} | sed 's#_1#_2#')
 # ... many lines of complex operations
done

Simplified version
for sample in ${DATA_DIR}/experimental_data/*_1.fastq.gz; do
 sample_name=$(basename ${sample} _1.fastq.gz)

 # Process paired-end reads
 fastp --in1 ${sample} \
 --in2 ${DATA_DIR}/experimental_data/${sample_name}_2.fastq.gz \
 --out1 ${RESULTS_DIR}/fastp/${sample_name}_1.clean.fastq.gz \
 --out2 ${RESULTS_DIR}/fastp/${sample_name}_2.clean.fastq.gz \
 --html ${RESULTS_DIR}/fastp/${sample_name}_report.html \
 --thread ${CPU_CORES}
done
```

```

Détails: : Variables

home > ubuntu > workspace > rnaseq > selected_notebooks > Pipe_03-variables.md

```
1 # Minimum Required Variables for Pipe_03
2
3 ## Core System Variables
4 1. `USER`
5     - Purpose: Student's username on Plasmabio
6     - Value: Automatically set by system
7     - Usage: Path construction for working directory
8     - Example: `${USER}`
9
10 ## Essential Working Directory Variables
11 1. `WORK_DIR`
12     - Purpose: Base working directory for bash analysis
13     - Value: `/srv/home/${USER}/meg_m2_rnaseq_bash`
14     - Usage: All student work and results storage
15     - Note: Created automatically at notebook startup
16
17 2. `DATA_DIR`
18     - Purpose: Location of shared RNA-seq data
19     - Value: `/srv/data/meg-m2-rnaseq`
20     - Usage: Access to experimental data
21     - Subpaths:
22         - `/srv/data/meg-m2-rnaseq/experimental_data/`
23
24 ## Results Directory Structure
25 1. `RESULTS_DIR`
26     - Purpose: Store all analysis outputs
27     - Value: ${WORK_DIR}/Results
28     - Subfolders:
29         - `fastp/`: Cleaned reads and FastP reports
30         - `fastq_screen/`: Contamination check results
31         - `multiqc/`: Combined quality reports
32
33 ## Processing Parameters
34 1. `CPU_CORES`
35     - Purpose: Control parallel processing
36     - Value: 4 (fixed for controlled environment)
37     - Usage: FastP and FastQ Screen threading
38     - Note: Simplified from authorizedCPU
```

```
## Simplified Variable Usage Example
```
Core directory setup
WORK_DIR="/srv/home/${USER}/meg_m2_rnaseq_bash"
DATA_DIR="/srv/data/meg-m2-rnaseq"
RESULTS_DIR="${WORK_DIR}/Results"
CPU_CORES=4

Create working directories
mkdir -p "${RESULTS_DIR}"/{fastp,fastq_screen,multiqc}

Process reads with fastp
fastp --in1 "${DATA_DIR}/experimental_data/sample_1.fastq.gz" \
 --in2 "${DATA_DIR}/experimental_data/sample_2.fastq.gz" \
 --out1 "${RESULTS_DIR}/fastp/sample_1.clean.fastq.gz" \
 --out2 "${RESULTS_DIR}/fastp/sample_2.clean.fastq.gz" \
 --html "${RESULTS_DIR}/fastp/sample_report.html" \
 --thread ${CPU_CORES}

Check contamination with FastQ Screen
fastq_screen --outdir "${RESULTS_DIR}/fastq_screen" \
 --threads ${CPU_CORES} \
 "${RESULTS_DIR}/fastp/*.clean.fastq.gz"

Generate MultiQC report
multiqc --outdir "${RESULTS_DIR}/multiqc" \
 --filename "quality_report" \
 "${RESULTS_DIR}"/{fastp,fastq_screen}
```
```

Détails: : Variables

```
## Variables Removed from Original
1. Removed IFB-specific variables:
  - `gohome` (replaced with `WORK_DIR`)
  - `rawfolder` (using `DATA_DIR`)
  - `fastpfolder` (using `RESULTS_DIR/fastp`)
  - `contascreenfolder` (using `RESULTS_DIR/fastq_screen`)
  - `qcsummaries` (using `RESULTS_DIR/multiqc`)
  - `logfolder` (logs stored in results directory)

2. Removed configuration variables:
  - `authorizedRAM` (using default values)
  - `inamemyfile` (using standardized names)
  - `mytitle` (simplified report headers)
  - `mycomment` (optional metadata)
  - `fscreenconffile` (using default config)
```

Benefits of Simplification

1. Clearer Structure
 - Consistent path hierarchy
 - Intuitive variable names
 - Reduced path complexity
2. JupyterHub Integration
 - Compatible with controlled environment
 - User-specific working directories
 - Read-only data access
3. Educational Benefits
 - Fewer variables to track
 - Clear data organization
 - Focus on analysis rather than setup
4. Maintenance Benefits
 - Easier to update paths
 - Consistent naming scheme
 - Simplified troubleshooting