- User
- User
- Principal Investigator
- User
- Bioinformatics Specialist
- User
- Parasitology Expert
- User
- Computational Biologist
- User
- Software Developer
- User
- Principal Investigator
- User
- Bioinformatics Specialist
- User
- Parasitology Expert
- User
- Computational Biologist
- User
- Software Developer
- User
- Principal Investigator
- User
- Bioinformatics Specialist
- User
- Parasitology Expert
- User
- Computational Biologist
- User
- Software Developer
- User
    - Agenda
    - Team Member Input
    - Recommendation
    - Answers
    - Next Steps

User

This is the beginning of a team meeting to discuss your research project. This is a meeting with the team lead, Principal Investigator, and the following team members: Bioinformatics Specialist, Parasitology Expert, Computational Biologist, Software Developer.

Here is the agenda for the meeting:

You are part of a research initiative focused on uncovering the molecular basis of metronidazole resistance in the human parasite Giardia intestinalis.

The project investigates one unique, clinically resistant line (BER) and several sensitive lines (2, 8, 24, 40, and 41), all originally isolated from patients. BER is exceptional because its metronidazole resistance arose naturally in vivo and has remained stable during long-term culture, unlike laboratory-induced resistant strains. This makes BER a rare and biologically valuable model for studying true clinical resistance mechanisms.

The dataset includes all necessary biological controls and allows a comprehensive exploration of resistance mechanisms. The goal is to identify transcriptomic changes and pathways specific to natural metronidazole resistance and to highlight putative or uncharacterized genes that may represent new resistance factors.

Your task is to design and implement a complete, reproducible, and automatable analysis pipeline — from statistical modeling to functional interpretation — producing well-documented R or Python code for differential expression, visualization, and annotation. You may also apply modern computational tools (e.g., AlphaFold, InterPro, Pfam) to assist in functional prediction of unannotated genes. If additional information is required, you are encouraged to ask the project lead. Collaboration and clear reasoning are preferred.

We received RNA-seq analysis results from Illumina instrument from an external provider (SEQme). Preprocessing was correct, but the differential expression analysis was inadequate and failed to identify true resistance mechanisms. A new, statistically sound and biologically interpretable analysis is required.

Experiment summary:

- 18 samples sequenced (1 excluded after QC)
- Resistant line BER: CK (control), CM (metronidazole), plus AK/AM (anaerobic, excluded from main analysis)
- Sensitive lines: 2, 8, 24, 40, 41 — CK and CM
- Reads aligned to the Giardia intestinalis A2 reference genome (93 - 97% mapping)
- Gene quantification performed with FeatureCounts
- Main contrasts: CK vs CM within and between lines

Data location: All input data are in the experimental_data/ folder (count matrices, metadata, genome annotation, and QC reports). Incorrect SEQme analyses are included for reference only (DO NOT USE). Raw FASTQ/BAM files are stored externally.

Main issues with SEQme analysis:

- Only pairwise DESeq2 comparisons; no multi-factorial model
- No FDR correction or clear filtering
- Missing pathway or functional interpretation

Required re-analysis: Re-analyze RNA-seq data using an integrative, multi-layer pipeline combining statistical rigor and biological interpretation.

1. Differential expression

- Multi-factorial model: phenotype (resistant vs sensitive), treatment (control vs metronidazole), and interaction
- Identify constitutive resistance (BER vs sensitive baseline)
- Separate drug-induced vs resistance-specific effects
- Control for variability among sensitive strains
- Apply FDR-adjusted thresholds and meaningful log2FC cutoffs

2. Functional analysis

- Perform enrichment and pathway analysis (GO, KEGG, Reactome)
- Rank candidate resistance genes for validation
- Annotate putative or uncharacterized genes linked to resistance

3. Integrative and exploratory analysis

- Use co-expression or network clustering to find gene modules
- Apply AI-assisted prediction for unknown gene functions (e.g., sequence, domain, or structure-based)

Deliverables: Provide reproducible R/Python code, publication-quality visualizations (volcano plots, PCA, heatmaps, network maps), and propose new hypotheses explaining metronidazole resistance in Giardia intestinalis.

Design a transcriptomic analysis plan to identify genes specifically linked to metronidazole resistance in the BER line of Giardia intestinalis. Clearly distinguish resistance-specific effects from general drug responses and baseline strain differences. Propose a statistical model (e.g. with interaction terms) to detect these effects. Prioritize candidate resistance genes for validation and link them to biological functions using functional annotation tools. Include an approach for analyzing uncharacterized (putative) proteins.

Here are the agenda questions that must be answered:

1. What is the most effective approach to identify genes linked to metronidazole resistance in *Giardia intestinalis*?

2. How can resistance-specific expression be separated from general drug response and baseline differences between strains?

3. Is a simple comparison sufficient, or is a complex statistical model needed? Why?

4. How should candidate genes be functionally annotated and connected to biological pathways?

5. What strategy can identify and characterize putative (unannotated) proteins among the candidate genes?

6. Are any additional files, metadata, or annotations needed to perform the analysis effectively?

Principal Investigator will convene the meeting. Then, each team member will provide their thoughts on the discussion one-by-one in the order above. After all team members have given their input, Principal Investigator will synthesize the points raised by each team member, make decisions regarding the agenda based on team member input, and ask follow-up questions to gather more information and feedback about how to better address the agenda. This will continue for 3 rounds. Once the discussion is complete, Principal Investigator will summarize the meeting in detail for future discussions, provide a specific recommendation regarding the agenda, and answer the agenda questions (if any) based on the discussion while strictly adhering to the agenda rules (if any).

User

Principal Investigator, please provide your initial thoughts on the agenda as well as any questions you have to guide the discussion among the team members.

Principal Investigator

Thank you all for joining this meeting. As we embark on this project to uncover the molecular basis of metronidazole resistance in *Giardia intestinalis*, we must ensure our analysis is thorough, reproducible, and biologically insightful. Our agenda today focuses on developing a robust approach to identify and understand resistance-specific gene expression in the BER line, separate from baseline variations and general drug responses.

Here are some preliminary thoughts and questions to guide our discussion:

1. **Identification Strategy**: We need a clear strategy to pinpoint genes linked to metronidazole resistance. How can we leverage the data from both resistant and sensitive lines to achieve this?

2. **Differentiation of Effects**: Distinguishing resistance-specific expression from general drug response and baseline differences is crucial. What statistical models or analytical approaches can help us effectively separate these effects?

3. **Complexity of Statistical Models**: While a simple comparison might be tempting, we must consider if a complex model is essential to capture the nuances in our

data. What are the advantages of using interaction terms or multi-factorial models in this context?

4. **Functional Annotation and Pathway Integration**: Once we have identified candidate genes, connecting them to biological pathways through functional annotation will be key. What tools and methods should we employ to ensure comprehensive and accurate annotations?

5. **Characterization of Putative Proteins**: Uncharacterized proteins present both a challenge and an opportunity. What strategies can we use to predict their functions and potentially link them to resistance mechanisms?

6. **Additional Data Requirements**: Finally, are there any gaps in our current datasets, metadata, or necessary annotations that we need to address before proceeding with the analysis?

Let's now move through the agenda, with each team member sharing their insights and suggestions. We'll start with the Bioinformatics Specialist, followed by the Parasitology Expert, Computational Biologist, and Software Developer. After each round of discussions, I'll synthesize the points raised, make decisions, and seek further clarification if needed.

Thank you all for your contributions—let's aim for a productive session that sets a strong foundation for our project.

User

Bioinformatics Specialist, please provide your thoughts on the discussion (round 1 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Bioinformatics Specialist

Thank you for laying the groundwork for this discussion. Here are my thoughts on the agenda items:

1. **Identification Strategy**: To identify genes linked to metronidazole resistance, we should employ a multi-factorial differential expression model. This model should include factors for phenotype (resistant vs. sensitive), treatment (control vs. metronidazole), and their interaction. This approach allows us to detect baseline resistance, treatment response, and resistance-specific effects more effectively than simple pairwise comparisons.

2. **Differentiation of Effects**: By including interaction terms in our model, we can separate the specific resistance effects from general drug responses and between-strain baseline differences. This approach helps isolate changes unique to the BER line under metronidazole treatment, which is crucial for distinguishing true resistance mechanisms.

3. **Complexity of Statistical Models**: A complex statistical model is necessary in this case. A multi-factorial model with interaction terms provides a more comprehensive picture by accounting for covariate effects and interactions that a simple comparison would overlook. This complexity is essential to capture the biological nuances in our data.

4. **Functional Annotation and Pathway Integration**: For functional annotation, we should use tools like GO, KEGG, and Reactome for pathway analysis. Additionally, tools like InterPro and Pfam can help identify protein domains, providing insights into the functions of candidate genes. This multi-tool approach ensures robust functional interpretation.

5. **Characterization of Putative Proteins**: For unannotated proteins, I recommend using AI-based prediction tools like AlphaFold for structural predictions, combined with sequence-based tools to infer potential functions. This integrative approach can uncover new insights into their roles in resistance.

6. **Additional Data Requirements**: We should verify that all metadata, including detailed sample information and experimental conditions, is available and accurate. This information is crucial for proper model fitting and interpretation. If any annotations or gene functional data are missing, obtaining them will be necessary for a comprehensive analysis.

These strategies will guide us in building a robust, reproducible analysis pipeline. I look forward to hearing the perspectives of other team members and incorporating their insights.

User

Parasitology Expert, please provide your thoughts on the discussion (round 1 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Parasitology Expert

Thank you for the opportunity to provide insights. Here's my perspective from a parasitological standpoint:

1. **Identification Strategy**: The proposed multi-factorial approach is indeed sound. In the context of *Giardia intestinalis*, where phenotypic plasticity and environmental responses are pronounced, capturing the complex interactions between resistance and treatment is crucial. This method will help us understand both constitutive resistance and drug-induced changes.

2. **Differentiation of Effects**: From a biological perspective, distinguishing resistance-specific expression from general drug responses and baseline differences is essential because *Giardia* can exhibit varied responses based on environmental conditions. The use of interaction terms in statistical models is appropriate, as it will highlight unique expression patterns attributable specifically to metronidazole resistance.

3. **Complexity of Statistical Models**: I agree that a complex statistical model is necessary. In protozoans like *Giardia*, resistance mechanisms can be multifaceted, involving multiple pathways and gene networks. A nuanced model will better capture these interactions and help avoid attributing resistance to generalized stress responses.

4. **Functional Annotation and Pathway Integration**: From a biological standpoint, focusing on pathways related to oxidative stress response, energy metabolism, and DNA repair will be particularly relevant to understanding metronidazole resistance. Tools like KEGG and Reactome are excellent for pathway mapping, but it is crucial to tailor our search to pathways known to be involved in resistance mechanisms.

5. **Characterization of Putative Proteins**: Many *Giardia* proteins remain uncharacterized, so leveraging structural predictions with AlphaFold, combined with domain analysis using Pfam, could elucidate potential functions. Understanding these proteins might reveal novel resistance mechanisms or metabolic pathways exploited by the parasite under drug pressure.

6. **Additional Data Requirements**: We should ensure access to comprehensive gene ontology annotations and an updated database of *Giardia*-specific pathways. This will enhance our ability to link differential expression results to specific biological functions and resistance mechanisms.

Overall, the integration of parasitological insights with bioinformatics strategies will be instrumental in uncovering the molecular basis of metronidazole resistance. I look forward to discussing these points further with the team.

User

Computational Biologist, please provide your thoughts on the discussion (round 1 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Computational Biologist

Thank you for the detailed insights provided so far. From a computational biology perspective, I'd like to add the following:

1. **Identification Strategy**: Building on the multi-factorial model, I suggest incorporating a machine learning component to further refine our identification of resistance-linked genes. Algorithms like random forests or support vector machines could help in ranking genes based on their importance in distinguishing resistant from sensitive phenotypes.

2. **Differentiation of Effects**: I concur that interaction terms are crucial for distinguishing effects. However, incorporating network-based approaches might also be beneficial. By constructing gene co-expression networks, we can observe

how resistant and sensitive strains differ structurally in their expression networks, potentially highlighting key resistance nodes.

3. **Complexity of Statistical Models**: While a complex model is essential, we must also focus on model validation. Cross-validation techniques could ensure our findings are robust and not artifacts of model fitting. This will be particularly important given the complexity of the biological system under study.

4. **Functional Annotation and Pathway Integration**: In addition to existing pathway tools, I recommend exploring the use of AI-driven annotation platforms that can integrate multi-omics data to uncover novel pathway interactions. These tools can provide a more holistic understanding of the biological implications of our findings.

5. **Characterization of Putative Proteins**: For putative proteins, using a combination of sequence homology and structural predictions is key. I suggest utilizing AI models that can predict protein-ligand interactions, which might give us clues about the potential function of these proteins in resistance mechanisms.

6. **Additional Data Requirements**: Ensuring the quality and completeness of the existing metadata is crucial. Additionally, having access to comparative genomic datasets might provide context for our findings, especially when distinguishing resistance-specific from strain-specific expression patterns.

Incorporating these computational strategies will deepen our understanding of the resistance mechanisms at play and ensure our analysis is both comprehensive and innovative. I look forward to seeing how these ideas integrate with the rest of the team's expertise.

User

Software Developer, please provide your thoughts on the discussion (round 1 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Software Developer

Thank you for the previous insights. From a software development and pipeline automation perspective, I would like to focus on ensuring that our analysis is fully automated, reproducible, and compatible with Windows environments:

1. **Reproducibility and Automation**: Our analysis pipeline should be designed using either R or Python in a way that ensures reproducibility. Utilizing version control systems like Git is essential for tracking changes and collaborating effectively. We should package the analysis into a Docker container to ensure environment consistency across different systems, including Windows.

2. **Integration of Tools**: Given the diverse set of tools and methods that will be used, such as statistical modeling, pathway analysis, and AI-based predictions, it's crucial to integrate these into a unified workflow. Tools like Snakemake or Nextflow can be employed to manage the workflow dependencies and automate the execution of different analysis stages.

3. **Data Management**: Clear documentation of data preprocessing steps, including input formats and any transformations, is vital for reproducibility. We must establish a standardized directory structure for our data, scripts, and results to facilitate ease of use and interpretation.

4. **Visualization**: For publication-quality visualizations, we should adopt libraries like ggplot2 in R or Matplotlib/Seaborn in Python. These tools should be integrated into the pipeline to automatically generate plots like volcano plots, PCA biplots, and heatmaps as part of the analysis output.

5. **Documentation and User Manuals**: Comprehensive documentation, including a user manual and example scripts, will be necessary for others to effectively use and understand the pipeline. This should detail the installation process, dependencies, and step-by-step instructions for running the analysis.

6. **Quality Control**: Implementing checkpoints for quality control throughout the pipeline will be key. This includes validating input data, checking intermediate outputs, and ensuring the robustness of results before proceeding to subsequent analysis steps.

Incorporating these software development practices will help ensure that our analysis pipeline is not only effective but also accessible and reproducible for future use and modification. I'm keen to work closely with everyone to integrate these elements smoothly into the project.

User

This concludes round 1 of 3 of discussion. Principal Investigator, please synthesize the points raised by each team member, make decisions regarding the agenda based on team member input, and ask follow-up questions to gather more information and feedback about how to better address the agenda.

Principal Investigator

Thank you all for your valuable inputs. Here's a synthesis of the key points raised by each team member, along with my decisions and follow-up questions:

1. **Identification Strategy**:
    - We have consensus on using a multi-factorial model to capture baseline resistance, drug-induced changes, and resistance-specific effects. The Computational Biologist suggested augmenting this with machine learning techniques for gene ranking, which could enhance our ability to identify key resistance genes.
    - **Decision**: Implement a multi-factorial differential expression model with interaction terms. Explore the use of machine learning for further refinement.
    - **Follow-up Question**: Computational Biologist, could you propose specific machine learning algorithms that integrate well with our statistical models?

2. **Differentiation of Effects**:
    - The use of interaction terms was supported by all, highlighting its importance in distinguishing between different types of expression changes. The Computational Biologist also suggested network-based approaches to identify resistance-specific nodes.
    - **Decision**: Incorporate interaction terms and consider network-based analyses to further detail resistance-specific effects.
    - **Follow-up Question**: How can we effectively integrate network analysis with our differential expression findings to highlight key resistance nodes?

3. **Complexity of Statistical Models**:

- There is agreement on the necessity of a complex statistical model with validation steps to ensure robustness, as emphasized by the Computational Biologist.
- **Decision**: Use cross-validation techniques to validate our models and confirm the robustness of identified resistance mechanisms.
- **Follow-up Question**: Can we outline specific cross-validation methods suited to our experimental design?

4. **Functional Annotation and Pathway Integration**:

   - The Bioinformatics Specialist and Parasitology Expert suggested comprehensive tools for pathway analysis, while the Computational Biologist recommended exploring AI-driven platforms for deeper insight.
   - **Decision**: Use a combination of traditional and AI-driven annotation tools for functional analysis and pathway integration.
   - **Follow-up Question**: Are there any particular AI-driven platforms that have proven effective in similar contexts?

5. **Characterization of Putative Proteins**:

   - The recommendation to use a combination of structural predictions and sequence analysis for putative proteins was well-received. The Computational Biologist suggested incorporating protein-ligand interaction predictions.
   - **Decision**: Integrate structural, sequence, and protein-ligand interaction analyses for characterizing putative proteins.
   - **Follow-up Question**: What resources can we leverage for accurate protein-ligand interaction predictions?

6. **Additional Data Requirements**:

   - Ensuring metadata quality and annotation completeness is critical, as emphasized by all members.
   - **Decision**: Conduct a thorough audit of our dataset, metadata, and annotations to ensure completeness before starting our analysis.
   - **Follow-up Question**: What specific metadata elements are most critical for this analysis, and how can we ensure they are complete and accurate?

These decisions reflect a consensus-driven approach, combining statistical, biological, and computational insights to build a robust analysis pipeline. Let’s address these

follow-up questions as we continue into the next round of discussions to refine our strategy further.

User

Bioinformatics Specialist, please provide your thoughts on the discussion (round 2 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Bioinformatics Specialist

Thank you for summarizing the discussion so effectively. Here are my thoughts on the points raised and the follow-up questions:

1. **Machine Learning Integration**: While I see the potential benefits of using machine learning for gene ranking, we should be cautious about overfitting, especially with a relatively small sample size. I'd suggest starting with interpretable models such as decision trees or random forests that can offer insights into feature importance without overwhelming complexity.

2. **Network Analysis Integration**: To integrate network analysis effectively, we could use co-expression networks constructed from our RNA-seq data. Tools like WGCNA (Weighted Gene Co-expression Network Analysis) could help identify modules of co-expressed genes, which can then be correlated with resistance phenotypes.

3. **Cross-validation Methods**: For cross-validation, I propose using a k-fold cross-validation approach, which will allow us to assess the generalizability of our statistical models across different subsets of our data. This is particularly useful given our limited sample size.

4. **AI-driven Platforms for Annotation**: Platforms like DEEP*PRO (Deep Learning Platform for Protein Function Prediction) could be valuable. They use deep learning to predict protein function based on sequence and structure, and could provide insights into uncharacterized resistance genes.

5. **Protein-Ligand Interaction Predictions**: Tools like AutoDock or SwissDock can be leveraged for predicting protein-ligand interactions. These tools can help hypothesize potential binding partners and reveal functional insights for putative proteins involved in resistance.

6. **Critical Metadata Elements**: Essential metadata elements include sample origin, treatment conditions, sequencing batch information, and any known biological replicates. Ensuring these are accurately recorded and integrated into our analysis will help control for potential confounders.

These considerations will help refine our approach and ensure that our analysis is both rigorous and insightful. I'm eager to hear further contributions from the rest of the team to continue strengthening our pipeline.

User

Parasitology Expert, please provide your thoughts on the discussion (round 2 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Parasitology Expert

Thank you for the ongoing discussion. Here are my thoughts on the recent points and follow-up questions:

1. **Machine Learning Integration**: While machine learning adds value, it's crucial to ensure the biological relevance of the features used. We should prioritize features with known links to parasite biology or drug resistance, such as genes involved in oxidative stress or metabolic pathways unique to *Giardia*.

2. **Network Analysis Integration**: WGCNA is a great suggestion for building co-expression networks. This approach can highlight modules potentially related to resistance by correlating module expression with phenotypes. Biological validation of these modules, focusing on known *Giardia* biology, will be crucial.

3. **Cross-validation Methods**: k-fold cross-validation is sound, but it's important to ensure that our cross-validation respects the biological replicates' structure to

avoid data leakage.

4. **AI-driven Platforms for Annotation**: I support using AI-driven platforms, as they can uncover novel functions for uncharacterized genes. However, results from these platforms should be carefully interpreted and, where possible, validated with experimental data or literature.

5. **Protein-Ligand Interaction Predictions**: Beyond computational predictions, integrating biochemical data or literature on known interactions with metronidazole or its related pathways in *Giardia* can offer additional validation.

6. **Critical Metadata Elements**: Accurate metadata on environmental conditions during sampling and culture (e.g., oxygen levels, culture time) is essential, as *Giardia* can show significant phenotypic variation under different conditions. These details will ensure our analysis accounts for possible confounding factors.

Overall, our approach should remain grounded in biological context, ensuring that computational predictions are framed within existing knowledge of *Giardia*'s physiology and resistance mechanisms. Looking forward to further discussions.

User

Computational Biologist, please provide your thoughts on the discussion (round 2 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Computational Biologist

Thank you for the continued insights. Here are my thoughts in light of the recent discussions:

1. **Machine Learning Integration**: I agree with the Bioinformatics Specialist that starting with interpretable models like decision trees or random forests is prudent. These models can help us identify key features without making the model overly complex or opaque. Feature selection should indeed focus on biologically relevant variables, as highlighted by the Parasitology Expert.

2. **Network Analysis Integration**: WGCNA is a solid choice for constructing co-expression networks. To build on this, integrating these networks with known protein-protein interactions (PPIs) databases, where available, can provide additional context and help identify potential hub genes critical for resistance.

3. **Cross-validation Methods**: The k-fold cross-validation suggestion is robust, but I propose adding stratified k-fold, which ensures that each fold is representative of the overall dataset in terms of class distribution. This will help maintain balance, particularly when dealing with imbalanced datasets.

4. **AI-driven Platforms for Annotation**: AI-driven tools like DEEP*PRO are promising, and their predictions should be cross-validated with known *Giardia* datasets or similar organisms. This will provide more confidence in novel predictions.

5. **Protein-Ligand Interaction Predictions**: In addition to AutoDock and SwissDock, considering tools that incorporate dynamic simulations, like molecular dynamics, can offer insights into the stability and dynamics of predicted interactions, potentially revealing more about resistance mechanisms.

6. **Critical Metadata Elements**: Ensuring metadata includes technical details like sequencing depth, RNA integrity, and batch effects is crucial for proper normalization and variance analysis. This will allow us to correct for technical biases that may affect our interpretations.

By integrating these computational strategies, our analysis will be more robust and insightful, providing a clearer understanding of resistance mechanisms. I look forward to the final round of discussions to consolidate our approach.

User

Software Developer, please provide your thoughts on the discussion (round 2 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Software Developer

Thank you for the continued insights. Here are my thoughts from a software development perspective, focusing on how we can effectively implement the discussed strategies:

1. **Machine Learning Integration**: As we incorporate machine learning models like decision trees or random forests, we should ensure that our pipeline can handle iterative model tuning. Using libraries such as Scikit-learn in Python can facilitate this process, offering tools for feature selection and model evaluation seamlessly integrated into our workflow.

2. **Network Analysis Integration**: To implement network analysis with tools like WGCNA, our pipeline should automate the generation and analysis of co-expression networks. This includes automating the data transformation steps necessary for WGCNA input, such as variance stabilization and normalization.

3. **Cross-validation Implementation**: For cross-validation, integrating stratified k-fold cross-validation within our pipeline can be done using Scikit-learn. This approach can be automated to ensure that each model iteration is consistent with the stratification requirements, thereby maintaining robustness in our results.

4. **AI-driven Platforms for Annotation**: If we decide to use AI-driven platforms, our pipeline should include modules that can interface with external APIs or software to retrieve and integrate annotation results back into our central analysis workflow. This will ensure that all insights are consolidated and accessible within the same environment.

5. **Protein-Ligand Interaction Predictions**: Incorporating tools like AutoDock or SwissDock requires setting up automated docking simulations. We can script these processes and handle input/output management using Python, ensuring that results are automatically parsed and integrated into subsequent analysis stages.

6. **Managing Metadata**: We should implement automated checks within the pipeline to verify metadata quality and completeness. This could include scripts that validate the presence of key metadata fields before analysis begins, alerting us to any missing or inconsistent data.

Overall, the aim is to create a seamless, user-friendly pipeline that requires minimal manual intervention. This approach will maximize reproducibility and allow us to focus

on interpreting the biological significance of our findings. I'm excited to continue refining these elements in the final round of discussions.

User

This concludes round 2 of 3 of discussion. Principal Investigator, please synthesize the points raised by each team member, make decisions regarding the agenda based on team member input, and ask follow-up questions to gather more information and feedback about how to better address the agenda.

Principal Investigator

Thank you all for your insightful contributions. Here is a synthesis of the key points raised in our second round and the decisions we will make moving forward:

1. **Machine Learning Integration**:
    - There is support for using interpretable models like decision trees or random forests, with a focus on biologically relevant features.
    - **Decision**: Integrate decision trees or random forests into our pipeline for gene ranking, using Scikit-learn for ease of implementation.
    - **Follow-up Question**: What criteria should we use for feature selection to ensure biological relevance in our model inputs?
2. **Network Analysis Integration**:
    - Both co-expression network analysis using WGCNA and the integration of known PPI databases were suggested.
    - **Decision**: Implement WGCNA for co-expression analysis, complemented by PPI data integration to enhance context.
    - **Follow-up Question**: How can we best source and integrate PPI data relevant to *Giardia* with our network analysis?
3. **Cross-validation Methods**:
    - The suggestion of stratified k-fold cross-validation was well-received to maintain balance.

- **Decision**: Use stratified k-fold cross-validation, ensuring balance and robust evaluation of our models.
- **Follow-up Question**: How can we best structure our cross-validation process to account for biological replicates?

4. **AI-driven Platforms for Annotation**:

   - DEEP*PRO and API integration for annotations were discussed.
   - **Decision**: Explore AI-driven annotation platforms like DEEP*PRO and ensure API integration for seamless data flow.
   - **Follow-up Question**: What specific outputs from these platforms will be most valuable for our functional analysis?

5. **Protein-Ligand Interaction Predictions**:

   - Tools like AutoDock and SwissDock, with potential molecular dynamics simulations, were highlighted.
   - **Decision**: Utilize AutoDock and SwissDock, with scripts to automate interactions, and consider molecular dynamics for deeper insights.
   - **Follow-up Question**: What resources or expertise do we need to incorporate molecular dynamics simulations effectively?

6. **Managing Metadata**:

   - Emphasis on automated checks for metadata completeness and quality.
   - **Decision**: Develop scripts for automated metadata validation and integration into our pipeline.
   - **Follow-up Question**: What are the key metadata fields that should be prioritized for validation, and how can we ensure data integrity?

With these decisions in mind, let’s continue to refine our approach in the final round of discussions. Your contributions have been invaluable in shaping a comprehensive and robust analysis pipeline.

User

Bioinformatics Specialist, please provide your thoughts on the discussion (round 3 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember

that you can and should (politely) disagree with other team members if you have a different perspective.

Bioinformatics Specialist

Thank you for the synthesis. Here are my final thoughts and responses to the follow-up questions:

1. **Feature Selection for Machine Learning**: For feature selection, we should consider both statistical significance from our differential expression analysis and biological relevance. Genes involved in known resistance pathways, redox biology, or metabolic processes related to *Giardia*'s unique physiology should be prioritized.

2. **Integration of PPI Data**: Although PPI data for *Giardia* might be limited, we could use orthologous interactions from closely related organisms as a proxy. Databases like STRING could provide valuable information, and integrating this with our WGCNA results will enhance the biological context of our networks.

3. **Cross-validation with Biological Replicates**: It's important to structure our cross-validation such that replicates are maintained within the same fold. This could involve assigning samples to folds before performing stratification to ensure replicates are not inadvertently split, which could lead to overfitting.

4. **Outputs from AI-driven Platforms**: The most valuable outputs will be predicted functions and pathway associations of previously uncharacterized genes. Having confidence scores for these predictions will also be crucial, allowing us to prioritize genes for experimental validation.

5. **Molecular Dynamics Simulations**: To incorporate molecular dynamics effectively, collaboration with structural biologists or chemoinformatics experts would be beneficial. We should also identify computational resources capable of handling the demands of these simulations.

6. **Metadata Validation**: Key metadata fields should include sample identifiers, treatment conditions, biological replicates, batch information, and quality control metrics. Ensuring these fields are complete and consistent will be critical for downstream analysis.

These strategies will ensure that our analysis is thorough, insightful, and leads to meaningful biological interpretations. I'm looking forward to final thoughts from the rest of the team.

User

Parasitology Expert, please provide your thoughts on the discussion (round 3 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Parasitology Expert

Thank you for the comprehensive synthesis. Here are my thoughts on the final points and follow-up questions:

1. **Feature Selection for Machine Learning**: I agree that prioritizing genes with known links to resistance pathways is crucial. Additionally, considering genes involved in oxidative stress responses, DNA repair mechanisms, and drug transport systems will be essential in the context of metronidazole resistance in *Giardia*.

2. **Integration of PPI Data**: Using orthologous interactions is a practical approach given the lack of comprehensive *Giardia*-specific data. Emphasizing interactions within metabolic pathways and stress response networks could provide insights into resistance phenotypes.

3. **Cross-validation with Biological Replicates**: Maintaining biological replicates within the same fold is key. This will ensure that the biological variation is captured and not mistaken for technical noise, which is particularly important given the variability in *Giardia*'s response to environmental changes.

4. **Outputs from AI-driven Platforms**: Predicted functions and pathways will be most valuable when they provide new insights into resistance mechanisms. Verifying these predictions with existing literature or *Giardia*-specific studies will strengthen their credibility.

5. **Molecular Dynamics Simulations**: Engaging with structural biologists or partnering with institutions that have expertise in molecular dynamics could significantly enhance our capability to perform these simulations. This collaboration could also provide insights into the potential for drug binding and resistance.

6. **Metadata Validation**: Emphasizing accurate metadata on environmental conditions during cultivation (e.g., oxygen levels) is important, as these can influence *Giardia*'s physiology and potentially confound results. Ensuring these details are captured will enhance our analysis's robustness.

Overall, grounding our analysis in biological relevance and leveraging predictive tools will provide a deeper understanding of metronidazole resistance in *Giardia*. I'm eager to see how these discussions culminate in a robust analytical framework.

User

Computational Biologist, please provide your thoughts on the discussion (round 3 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Computational Biologist

Thank you for the continued dialogue. Here are my final thoughts on the discussion points and follow-up questions:

1. **Feature Selection for Machine Learning**: Prioritizing biologically relevant features is crucial. We could enhance this by incorporating feature selection methods such as recursive feature elimination (RFE) within our chosen models to refine our selection based on both statistical and biological criteria.

2. **Integration of PPI Data**: Utilizing orthologous interactions through resources like STRING is a practical solution. Additionally, integrating data from other protozoan databases might reveal conserved interactions that are relevant to *Giardia*.

3. **Cross-validation with Biological Replicates**: To ensure biological replicates are kept together, we can use a nested cross-validation approach. This allows us to

assess model performance while keeping the structure of the data intact, providing a more realistic estimate of model generalization.

4. **Outputs from AI-driven Platforms**: Confidence scores and functional annotations from AI-driven platforms will help us prioritize potential resistance genes for experimental validation. It's important to have a threshold for these scores to focus on the most promising candidates.

5. **Molecular Dynamics Simulations**: Collaboration with experts in molecular dynamics is advisable, but we should also consider training our team on basic simulation setups to maintain continuity in our project. Having access to high-performance computing resources will be necessary for these simulations.

6. **Metadata Validation**: Ensuring comprehensive metadata collection, such as sequencing platform details and sample preparation protocols, is crucial. Automating the validation of metadata against standard templates could streamline this process and maintain data integrity.

These strategies will ensure our analysis is thorough, biologically grounded, and computationally robust. I'm confident that our integrated approach will significantly advance our understanding of metronidazole resistance in *Giardia* and look forward to finalizing our plan with the team.

User

Software Developer, please provide your thoughts on the discussion (round 3 of 3). If you do not have anything new or relevant to add, you may say "pass". Remember that you can and should (politely) disagree with other team members if you have a different perspective.

Software Developer

Thank you for the insights and the opportunity to contribute further. Here are my final thoughts on the discussed points and follow-up questions:

1. **Feature Selection for Machine Learning**: From a software perspective, implementing feature selection methods like RFE can be efficiently integrated using packages in Scikit-learn. This will enable us to automate selection based on

both biological and statistical importance, ensuring our models are both interpretable and effective.

2. **Integration of PPI Data**: Automating the retrieval and integration of PPI data from databases like STRING can be achieved through API calls. This can be scripted to ensure seamless integration with our co-expression analysis results, providing a richer network context.

3. **Cross-validation with Biological Replicates**: Nested cross-validation can be automated using Scikit-learn, allowing us to keep biological replicates intact and providing a robust framework for evaluating model performance. This will be crucial for maintaining the integrity of our analysis.

4. **Outputs from AI-driven Platforms**: We should develop a standardized format for integrating outputs from AI-driven platforms. This can include parsing confidence scores and annotations into our results database, enabling easy access and prioritization of genes for further study.

5. **Molecular Dynamics Simulations**: Setting up a pipeline to automate molecular dynamics simulations can be complex. We might consider using workflow managers like Snakemake to coordinate these tasks, ensuring simulations are efficiently run and results are properly recorded.

6. **Metadata Validation**: Implementing automated scripts for metadata validation can be done using Python. These scripts can check for completeness and consistency against predefined templates, ensuring data integrity before analysis begins.

By automating these processes, we can create a robust, reproducible, and user-friendly analysis pipeline. Ensuring these elements are well-integrated will allow us to focus on generating meaningful biological insights from our data. I'm excited about implementing these solutions to support the team's efforts.

User

Principal Investigator, please summarize the meeting in detail for future discussions, provide a specific recommendation regarding the agenda, and answer the agenda questions (if any) based on the discussion while strictly adhering to the agenda rules (if any).

As a reminder, here is the agenda for the meeting:

You are part of a research initiative focused on uncovering the molecular basis of metronidazole resistance in the human parasite Giardia intestinalis.

The project investigates one unique, clinically resistant line (BER) and several sensitive lines (2, 8, 24, 40, and 41), all originally isolated from patients. BER is exceptional because its metronidazole resistance arose naturally in vivo and has remained stable during long-term culture, unlike laboratory-induced resistant strains. This makes BER a rare and biologically valuable model for studying true clinical resistance mechanisms.

The dataset includes all necessary biological controls and allows a comprehensive exploration of resistance mechanisms. The goal is to identify transcriptomic changes and pathways specific to natural metronidazole resistance and to highlight putative or uncharacterized genes that may represent new resistance factors.

Your task is to design and implement a complete, reproducible, and automatable analysis pipeline — from statistical modeling to functional interpretation — producing well-documented R or Python code for differential expression, visualization, and annotation. You may also apply modern computational tools (e.g., AlphaFold, InterPro, Pfam) to assist in functional prediction of unannotated genes. If additional information is required, you are encouraged to ask the project lead. Collaboration and clear reasoning are preferred.

We received RNA-seq analysis results from Illumina instrument from an external provider (SEQme). Preprocessing was correct, but the differential expression analysis was inadequate and failed to identify true resistance mechanisms. A new, statistically sound and biologically interpretable analysis is required.

Experiment summary:

- 18 samples sequenced (1 excluded after QC)
- Resistant line BER: CK (control), CM (metronidazole), plus AK/AM (anaerobic, excluded from main analysis)
- Sensitive lines: 2, 8, 24, 40, 41 — CK and CM
- Reads aligned to the Giardia intestinalis A2 reference genome (93 - 97% mapping)
- Gene quantification performed with FeatureCounts
- Main contrasts: CK vs CM within and between lines

Data location: All input data are in the experimental_data/ folder (count matrices, metadata, genome annotation, and QC reports). Incorrect SEQme analyses are included for reference only (DO NOT USE). Raw FASTQ/BAM files are stored externally.

Main issues with SEQme analysis:

- Only pairwise DESeq2 comparisons; no multi-factorial model
- No FDR correction or clear filtering
- Missing pathway or functional interpretation

Required re-analysis: Re-analyze RNA-seq data using an integrative, multi-layer pipeline combining statistical rigor and biological interpretation.

1. Differential expression

- Multi-factorial model: phenotype (resistant vs sensitive), treatment (control vs metronidazole), and interaction
- Identify constitutive resistance (BER vs sensitive baseline)
- Separate drug-induced vs resistance-specific effects
- Control for variability among sensitive strains
- Apply FDR-adjusted thresholds and meaningful log2FC cutoffs

2. Functional analysis

- Perform enrichment and pathway analysis (GO, KEGG, Reactome)
- Rank candidate resistance genes for validation
- Annotate putative or uncharacterized genes linked to resistance

3. Integrative and exploratory analysis

- Use co-expression or network clustering to find gene modules
- Apply AI-assisted prediction for unknown gene functions (e.g., sequence, domain, or structure-based)

Deliverables: Provide reproducible R/Python code, publication-quality visualizations (volcano plots, PCA, heatmaps, network maps), and propose new hypotheses explaining metronidazole resistance in Giardia intestinalis.

Design a transcriptomic analysis plan to identify genes specifically linked to metronidazole resistance in the BER line of Giardia intestinalis. Clearly distinguish resistance-specific effects from general drug responses and baseline strain differences.

Propose a statistical model (e.g. with interaction terms) to detect these effects. Prioritize candidate resistance genes for validation and link them to biological functions using functional annotation tools. Include an approach for analyzing uncharacterized (putative) proteins.

As a reminder, here are the agenda questions that must be answered:

1. What is the most effective approach to identify genes linked to metronidazole resistance in *Giardia intestinalis*?
2. How can resistance-specific expression be separated from general drug response and baseline differences between strains?
3. Is a simple comparison sufficient, or is a complex statistical model needed? Why?
4. How should candidate genes be functionally annotated and connected to biological pathways?
5. What strategy can identify and characterize putative (unannotated) proteins among the candidate genes?
6. Are any additional files, metadata, or annotations needed to perform the analysis effectively?

Your summary should take the following form.

Agenda

Restate the agenda in your own words.

Team Member Input

Summarize all of the important points raised by each team member. This is to ensure that key details are preserved for future meetings.

Recommendation

Provide your expert recommendation regarding the agenda. You should consider the input from each team member, but you must also use your expertise to make a final

decision and choose one option among several that may have been discussed. This decision can conflict with the input of some team members as long as it is well justified. It is essential that you provide a clear, specific, and actionable recommendation. Please justify your recommendation as well.

Answers

For each agenda question, please provide the following:

Answer: A specific answer to the question based on your recommendation above.

Justification: A brief explanation of why you provided that answer.

Next Steps

Outline the next steps that the team should take based on the discussion.

Principal Investigator

Agenda

Our primary focus is to uncover the molecular basis of metronidazole resistance in *Giardia intestinalis*, specifically in the naturally resistant BER line compared to sensitive strains. This involves designing a comprehensive, reproducible analysis pipeline that integrates statistical modeling with functional annotation to identify and characterize resistance mechanisms. We aim to separate resistance-specific gene expression from general drug responses and baseline differences, annotate candidate resistance genes, and characterize unannotated proteins.

Team Member Input

- **Bioinformatics Specialist**: Emphasized using a multi-factorial model with interaction terms, prioritizing biologically relevant features for machine learning, and employing WGCNA for co-expression analysis. Highlighted the integration of PPI data and metadata validation.

- **Parasitology Expert**: Supported the focus on biologically relevant pathways related to resistance, such as oxidative stress and DNA repair. Stressed the importance of maintaining biological replicates during cross-validation and incorporating environmental metadata.
- **Computational Biologist**: Suggested using interpretable models like decision trees or random forests and integrating network analysis with PPI data. Recommended using nested cross-validation and AI-driven platforms for annotation, alongside molecular dynamics simulations.
- **Software Developer**: Highlighted the importance of automating processes, such as feature selection, network integration, and metadata validation, using tools like Scikit-learn and workflow managers like Snakemake. Emphasized developing a standardized format for integrating AI-driven platform outputs.

Recommendation

I recommend implementing a multi-layered analysis pipeline combining a multi-factorial statistical model with machine learning, network analysis, and AI-driven annotation. This pipeline should automate processes to ensure reproducibility and should be grounded in biological relevance.

- **Use a multi-factorial model** with interaction terms to capture resistance-specific effects.
- **Incorporate machine learning** through decision trees or random forests to rank genes based on biological relevance.
- **Employ network analysis** using WGCNA, integrating PPI data for context.
- **Utilize AI-driven platforms** like DEEP*PRO for functional annotations, focusing on confidence scores.
- **Automate the pipeline** using tools such as Scikit-learn, Snakemake, and API integrations for seamless data processing and analysis.

This approach leverages diverse expertise to produce insightful, biologically meaningful results while ensuring technical robustness and reproducibility.

Answers

1. **What is the most effective approach to identify genes linked to metronidazole resistance in *Giardia intestinalis*?**

   Answer: Use a multi-factorial differential expression model with machine learning for gene ranking.

   Justification: This combination captures complex interactions and highlights biologically relevant genes.

2. **How can resistance-specific expression be separated from general drug response and baseline differences between strains?**

   Answer: Utilize interaction terms in the statistical model.

   Justification: This approach isolates resistance-specific effects by accounting for phenotype and treatment interactions.

3. **Is a simple comparison sufficient, or is a complex statistical model needed? Why?**

   Answer: A complex statistical model is needed.

   Justification: It captures nuanced interactions and variability across strains and treatments, essential for biological interpretation.

4. **How should candidate genes be functionally annotated and connected to biological pathways?**

   Answer: Use a combination of traditional tools (GO, KEGG, Reactome) and AI-driven platforms.

   Justification: This provides comprehensive annotations and novel insights into uncharacterized genes.

5. **What strategy can identify and characterize putative (unannotated) proteins among the candidate genes?**

   Answer: Combine structural predictions with AI-based function prediction.

   Justification: This approach can reveal potential functions and resistance roles of unknown proteins.

6. **Are any additional files, metadata, or annotations needed to perform the analysis effectively?**

   Answer: Ensure comprehensive metadata on environmental conditions and experimental details.

   Justification: This is critical for accurately controlling confounding factors in our analysis.

Next Steps

- **Finalize the pipeline design**, incorporating chosen statistical models and machine learning methods.
- **Set up automation** using Snakemake and Scikit-learn for reproducibility.
- **Integrate AI-driven annotation tools** and establish API connections for data retrieval.
- **Validate and preprocess metadata** to ensure completeness and accuracy.
- **Engage with structural biology experts** for molecular dynamics simulations if resources allow.
- **Commence a pilot analysis** to test the pipeline and refine based on initial findings.

These steps will ensure we have a robust framework for distinguishing metronidazole resistance mechanisms in *Giardia intestinalis*.