# Next Steps: Vitamin D and Type 2 Diabetes Research Project

## Project Status

✅ **Phase 1 Complete**: Literature review, hypothesis development, aims paper, and computational infrastructure setup

## Immediate Next Steps (Priority Order)

### 1. Data Acquisition and Access 🔐

**Timeline**: 1-2 months

#### Action Items:

- [ ] **Apply for dbGaP Access**
- Submit data access request for identified datasets (ARIC, JHS, CARDIA, WHI)
- Prepare Data Use Certification (DUC)
- Complete required training (e.g., "Protecting Human Research Participants")
- Estimated approval time: 4-8 weeks
- [ ] **Institutional Approvals**
- Obtain IRB approval/exemption for secondary data analysis
- Get institutional signing official approval for dbGaP
- Secure data storage infrastructure approval
- [ ] **UK Biobank Application**
- Submit research proposal to UK Biobank
- Request African ancestry subset data
- Budget for access fees (~$3,000-5,000)

#### Resources Needed:

- Institutional credentials and signing official
- Secure computing environment (meets NIH security requirements)
- Budget for UK Biobank access

---

### 2. Real Data Analysis Pipeline Development 🔬

**Timeline**: 2-3 months (parallel with data acquisition)

#### Genomics Analysis:

- [ ] **Quality Control Pipeline**
- Implement ancestry verification using PCA
- Set up genotype QC filters (call rate, HWE, MAF)

- Prepare imputation pipeline (TOPMed reference panel)
- [ ] **GWAS Analysis**
- Adapt existing scripts for real data
- Implement population stratification correction (PCs)
- Set up conditional analysis for independent signals
- Plan meta-analysis across cohorts
- [ ] **Functional Annotation**
- Set up VEP (Variant Effect Predictor)
- Integrate CADD scores, PolyPhen, SIFT
- Map variants to vitamin D pathway genes

### Proteomics Analysis:

- [ ] **Data Preprocessing**
- Normalize protein abundance data
- Handle missing values appropriately
- Batch effect correction
- [ ] **Association Testing**
- Protein-T2D associations
- Protein-vitamin D associations
- Mediation analysis framework

### Metabolomics Analysis:

- [ ] **Metabolite Profiling**
- Identify vitamin D metabolites
- Glucose metabolism markers
- Lipid profiles
- [ ] **Pathway Analysis**
- KEGG pathway enrichment
- Metabolite set enrichment analysis

---

## 3. Multi-Omics Integration 🧬

**Timeline**: 3-4 months

### Integration Approaches:

- [ ] **Mendelian Randomization**
- Implement two-sample MR
- Test vitamin D → T2D causality
- Sensitivity analyses (MR-Egger, weighted median)
- [ ] **Network Analysis**

- Build gene-protein-metabolite networks
- Identify key regulatory nodes
- Community detection algorithms

- [ ] **Machine Learning Models**

- Develop predictive models for T2D risk
- Feature importance analysis
- Cross-validation strategies

---

## 4. Manuscript Preparation 📝

**Timeline**: 4-6 months

### Primary Manuscript:

- [ ] **Results Section**
- Generate all figures and tables
- Write comprehensive results narrative
- Statistical validation

- [ ] **Discussion**

- Interpret findings in biological context
- Compare with existing literature
- Address limitations
- Clinical implications

- [ ] **Target Journals**

- Primary: Nature Genetics, Nature Communications
- Secondary: Diabetes, Diabetologia
- Backup: PLoS Genetics, BMC Genomics

### Supplementary Materials:

- [ ] Supplementary figures and tables
- [ ] Detailed methods
- [ ] Code availability (GitHub repository)
- [ ] Data availability statements

---

## 5. Thesis Committee Milestones 🎓

### Committee Meeting #1 (Month 3-4):

- [ ] Present data acquisition progress
- [ ] Show preliminary QC results
- [ ] Discuss any challenges with real data

### Committee Meeting #2 (Month 6-7):

- [ ] Present initial GWAS findings

- [ ] Show proteomics/metabolomics results
- [ ] Discuss integration strategies

### Committee Meeting #3 (Month 9-10):

- [ ] Present integrated multi-omics results
- [ ] Show draft manuscript figures
- [ ] Discuss publication timeline

### Thesis Defense (Month 12-15):

- [ ] Complete manuscript submission
- [ ] Prepare comprehensive thesis document
- [ ] Create defense presentation

---

## 6. Skills Development 📚

### Computational Skills:

- [ ] **Advanced R/Bioconductor**
- GWAS packages (PLINK, GCTA, BOLT-LMM)
- Proteomics (limma, DEqMS)
- Metabolomics (xcms, MetaboAnalystR)
- [ ] **Python for Bioinformatics**
- Pandas for data manipulation
- Scikit-learn for ML
- NetworkX for network analysis
- [ ] **High-Performance Computing**
- Cluster job submission (SLURM/PBS)
- Parallel processing
- Memory optimization

### Statistical Methods:

- [ ] Mendelian Randomization theory and practice
- [ ] Mixed models for related individuals
- [ ] Multiple testing correction strategies
- [ ] Causal inference methods

---

## 7. Collaboration and Networking 🤝

### Internal Collaborations:

- [ ] Identify statistical genetics expert for consultation
- [ ] Connect with proteomics/metabolomics core facilities
- [ ] Engage clinical collaborators for interpretation

### External Networking:

- [ ] Present at departmental seminars

- [ ] Submit abstracts to conferences:
- American Society of Human Genetics (ASHG)
- American Diabetes Association (ADA)
- Keystone Symposia
- [ ] Join relevant working groups (e.g., T2D-GENES, CHARGE)

---

## 8. Funding Opportunities 💰

### Predoctoral Fellowships:

- [ ] **NIH F31** (Ruth L. Kirschstein NRSA)
- Deadline: April, August, December
- Supports 2-3 years of PhD research
- [ ] **ADA Predoctoral Fellowship**
- Deadline: Usually January
- Diabetes-focused research
- [ ] **Diversity Supplements**
- If PI has active NIH grant
- Rolling deadlines

### Travel Grants:

- [ ] Conference-specific travel awards
- [ ] University graduate student travel funds
- [ ] Professional society student awards

---

# Risk Mitigation Strategies

## Potential Challenges:

1. **Data Access Delays**
   - **Mitigation**: Apply early, have backup datasets identified
   - **Alternative**: Use summary statistics for initial analyses

2. **Limited Sample Sizes**
   - **Mitigation**: Meta-analysis across multiple cohorts
   - **Alternative**: Focus on effect size rather than just significance

3. **Null Findings**
   - **Mitigation**: Frame as important negative results
   - **Alternative**: Emphasize methodological contributions

4. **Technical Challenges**
   - **Mitigation**: Build in buffer time for troubleshooting
   - **Alternative**: Seek expert consultation early

---

## Success Metrics

### Year 1:

- ✅ Aims paper complete
- ✅ Computational infrastructure ready
- 🎯 Data access obtained
- 🎯 Initial GWAS results

### Year 2:

- 🎯 Multi-omics integration complete
- 🎯 First manuscript submitted
- 🎯 Conference presentation

### Year 3:

- 🎯 Manuscript published
- 🎯 Thesis defense
- 🎯 PhD degree conferred

## Resources and Support

### Computational Resources:

- University HPC cluster
- Cloud computing credits (AWS, Google Cloud)
- Local workstation for development

### Data Storage:

- Secure server meeting NIH requirements
- Encrypted backup systems
- Version control (Git/GitHub)

### Mentorship:

- Primary advisor (weekly meetings)
- Thesis committee (quarterly meetings)
- Statistical genetics consultant (as needed)
- Clinical collaborator (monthly check-ins)

## Timeline Visualization

```
Month 1-2:   Data access applications
Month 2-4:   Pipeline development & testing
Month 4-6:   Real data QC & initial analyses
Month 6-9:   Multi-omics integration
Month 9-12:  Manuscript writing
Month 12-15: Revisions & thesis preparation
Month 15-18: Defense & graduation
```

---

## Contact and Collaboration

For questions or collaboration opportunities related to this project:

- **Repository**: https://github.com/ej777spirit/Abacus-VitD-DM2
- **Primary Investigator**: [Your Name]
- **Institution**: [Your Institution]

---

**Last Updated**: October 1, 2025
**Status**: Phase 1 Complete - Ready for Data Acquisition Phase