Publicly Available Omics Datasets Inventory

Vitamin D, Type 2 Diabetes, and African Ancestry Populations

Document Purpose: This inventory catalogs real, publicly available omics datasets for hierarchical multi-omics research on vitamin D and Type 2 diabetes in African ancestry populations.

Last Updated: September 30, 2025

Table of Contents

GENOMICS DATA (Priority 1)

1. Africa America Diabetes Mellitus (AADM) Study - Type 2 Diabetes GWAS

Dataset ID: phs001844.v1.p1
Repository: dbGaP (Database of Genotypes and Phenotypes)
Direct URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001844.v1.p1

Data Type:
- Whole genome SNP genotyping
- Genome-wide association study (GWAS) data
- Imputed genotypes

Platforms:
- Axiom™ PanAFR SNP array (n=1,808 samples)
- Multi-Ethnic Global Array (MEGA) (n=3,423 samples)
- **Total samples: ~5,231**

Population:
- **Sub-Saharan African populations**
- Countries: Nigeria, Ghana, Kenya
- Cases and controls for Type 2 diabetes

Sample Characteristics:
- Cases: Participants meeting ADA criteria for T2D (FPG ≥126 mg/dl or 2-hr glucose ≥200 mg/dl)

- Controls: FPG <110 mg/dl, 2-hr glucose <140 mg/dl
- Exclusions: Type 1 diabetes (GAD autoantibodies, low C-peptide)

Relevance to Research:
- ✓ Type 2 diabetes genetic associations
- ✓ African ancestry-specific variants
- ✓ TCF7L2 and ZRANB3 gene associations
- ✓ Imputation using African Genome Resources Haplotype Reference Panel
- Potential linkage to vitamin D pathway genes (VDR, GC, CYP27B1) through GWAS results

Data Available:
- Genotype data (imputed and raw)
- Phenotype data (clinical measurements, BMI, glucose levels)
- Quality-controlled SNP data (MAF ≥0.01, info score ≥0.3)
- Principal components for ancestry adjustment

Access Method:
- **Controlled access** - Requires dbGaP application
- Submit Data Access Request (DAR) through dbGaP
- IRB approval required
- Institutional signing official needed
- Processing time: 2-4 weeks typically

File Formats:
- PLINK format (.bed, .bim, .fam)
- VCF (Variant Call Format)
- Phenotype files (tab-delimited text)

Preprocessing Information:
- Quality control filters applied
- Imputation completed using African reference panel
- Population stratification adjustment (first 3 PCs)
- Genetic relatedness matrix computed

Principal Investigator: Charles Rotimi, PhD (NIH)
Funding: NIH (3T37TW00041-03S2, R01-DK54001, ZIAHG200362)

Download Instructions:
1. Create dbGaP account at https://dbgap.ncbi.nlm.nih.gov/
2. Complete Data Access Request for phs001844
3. Obtain IRB approval and institutional signatures
4. Upon approval, download via dbGaP FTP or Aspera
5. Use dbGaP download tools (SRA Toolkit for sequence data)

---

2. T2D-GENES Multi-Ethnic Follow-up Study - Whole Exome Sequencing

Dataset ID: phs001610.v6.p16
Repository: dbGaP
Direct URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001610.v6.p16

Data Type:
- Whole exome sequencing (WES)
- Next-generation sequencing data
- Variant calls and annotations

Platforms:
- Broad Institute: Agilent v2 capture, Illumina HiSeq
- Multiple sequencing centers across consortia

Population:
- **Multi-ethnic with substantial African American representation**
- **African American cohorts:**
- Jackson Heart Study: 500 cases, 526 controls (n=1,026)
- Wake Forest School of Medicine: 518 cases, 530 controls (n=1,048)
- ESP African Americans: 467 cases, 1,374 controls (n=1,841)
- BioMe Biobank: 1,297 cases, 1,256 controls (n=2,553)
- **Total African Americans: ~6,468 samples**
- Also includes: East Asian, South Asian, Hispanic, European cohorts
- **Overall study: ~52,000 samples across all ancestries**

Sample Characteristics:
- Cases: T2D diagnosis, medication use, non-fasting glucose >200 mg/dL, or diagnosis <35 years
- Controls: No T2D history
- Framingham Heart Study subset: 396 T2D cases + 596 controls = 992 samples

Relevance to Research:
- ✓ Type 2 diabetes rare variant associations
- ✓ Coding variants in T2D susceptibility genes
- ✓ African American-specific exome variants
- ✓ Multi-ethnic comparison for ancestry-specific effects
- Can be cross-referenced with vitamin D pathway genes

Data Available:
- Aligned sequence reads (BAM files)
- Variant calls (VCF format)
- Phenotype data (diabetes status, clinical measurements)
- Quality metrics
- Ancestry assignments

Access Method:
- **Controlled access** - dbGaP application required
- Separate DAR may be needed for each sub-cohort
- IRB approval required
- Data Use Limitations apply

File Formats:
- BAM (Binary Alignment Map)
- VCF (Variant Call Format)
- CRAM (compressed BAM)
- Phenotype files (tab-delimited)

Preprocessing Information:
- Sequence alignment to human reference genome

- Variant calling using GATK or similar pipelines
- Quality score filtering
- Annotation with functional consequences

Principal Investigators:
- Jose Florez (Broad Institute, Mass General)
- Michael Boehnke (University of Michigan)
- Mark McCarthy (Wellcome Trust, Oxford)
- David Altshuler (Broad Institute)

Funding: NIDDK (U01DK085526), NHGRI (U54HG003067)

Related Resources:
- T2D-GENES Consortium: https://t2dgenes.org
- Framingham Heart Study: https://www.framinghamheartstudy.org/

Download Instructions:
1. Access dbGaP and submit DAR for phs001610
2. Specify sub-studies of interest (Jackson Heart, Wake Forest, etc.)
3. Download using dbGaP repository tools
4. BAM files are large (50-100 GB per sample); use Aspera for faster transfer
5. VCF files more manageable for variant-level analysis

---

3. T2D-GENES Project 1: Ashkenazi (Includes African American Data)

Dataset ID: phs001095.v1.p1
Repository: dbGaP
Direct URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001095.v1.p1

Data Type:
- Exome sequencing
- Type 2 diabetes genetic variants

Platforms:
- Next-generation sequencing

Population:
- Primary: Ashkenazi Jewish (858 participants)
- **Also includes African American cohorts from broader T2D-GENES:**
- Jackson Heart Study
- Wake Forest School of Medicine Study

Sample Characteristics:
- T2D cases and controls
- Exome-level genetic variation

Relevance to Research:
- ✓ Type 2 diabetes coding variants
- ✓ Cross-ancestry comparison potential
- Can identify shared and ancestry-specific variants
- Part of larger T2D-GENES consortium

Access Method:
- **Controlled access** - dbGaP application
- May need separate applications for different cohorts

File Formats:
- VCF (variant calls)
- BAM (sequence alignments)
- Phenotype data files

Funding: Part of T2D-GENES consortium (NIDDK funding)

Download Instructions:
1. Submit dbGaP DAR for phs001095
2. Access through standard dbGaP protocols
3. Cross-reference with phs001610 for comprehensive T2D-GENES data

---

4. African American Hepatocyte Gene Expression and Admixture

Dataset ID: GSE124076
Repository: NCBI Gene Expression Omnibus (GEO)
Direct URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124076

Data Type:
- RNA-seq (gene expression profiling)
- DNA methylation profiling (Illumina 850K array)
- SNP genotyping (genome variation)

Platforms:
- **GPL16791:** Illumina HiSeq 2500
- **GPL20301:** Illumina HiSeq 4000
- **GPL23976:** Illumina Infinium HumanMethylation850 BeadChip
- **GPL24127:** Illumina Infinium Multi-Ethnic Global-8 v1.0 Array

Population:
- **African American hepatocytes**
- Sample size: 567 samples (includes various treatment conditions)
- Individuals with varying proportions of African ancestry

Sample Characteristics:
- Primary human hepatocytes
- Various drug treatment conditions:
- Omeprazole-treated
- Phenobarbital-treated
- Dexamethasone-treated
- Carbamazepine-treated
- Phenytoin-treated
- Control (untreated)
- Multiple biological replicates per condition

Relevance to Research:
- ✓ **VDR gene expression in African American hepatocytes**
- ✓ Vitamin D metabolism genes (CYP27B1, CYP24A1, GC)

- ✓ Association with West African ancestry
- ✓ Gene expression × ancestry interactions
- ✓ Epigenetic regulation (DNA methylation)
- Hepatic insulin signaling pathways
- Drug metabolism and response (including vitamin D metabolism)

Data Available:
- RNA-seq read counts and TPM values
- Differential gene expression analysis results
- DNA methylation beta values (850K sites)
- SNP genotypes for ancestry estimation
- Admixture proportions (West African ancestry)
- Sample metadata (treatment, ancestry estimates)

Access Method:
- **Public access** - No application required
- Direct download from GEO

File Formats:
- RNA-seq: FASTQ (raw reads), BAM (aligned), TXT (count matrices)
- Methylation: IDAT files, TXT (beta values)
- Genotypes: VCF or PLINK format
- Series Matrix: Tab-delimited metadata and processed data

Preprocessing Information:
- RNA-seq aligned to human reference genome
- Gene expression quantified using standard pipelines
- Methylation data processed with minfi or similar
- Quality control and normalization applied
- Ancestry estimated from genome-wide SNPs

Principal Investigator: Minoli Perera (Northwestern University)
NIH Grant: R01 MD009217 (Health disparity in pharmacogenomics: African American SNPs and drug metabolism)

Publication: PMID 31798965

SubSeries Components:
- **GSE123995:** Methylation data
- **GSE124074:** RNA-seq data
- **GSE147628:** African American hepatocyte genotype
- **GSE222593:** Additional RNA-seq

Download Instructions:
1. Visit GEO accession page: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124076
2. Click “Download family” for full dataset
3. Or download individual files:
- **GSE124076_RAW.tar** (3.3 GB) - raw IDAT and TXT files
- Series Matrix files for processed data
- Supplementary files with metadata
4. Use GEOquery R package for programmatic access:
```
R
  library(GEOquery)
```

```
gse <- getGEO("GSE124076")
```

5. Access raw sequencing data via SRA (BioProject: PRJNA510661)

Analysis Notes:
- VDR expression can be extracted from RNA-seq data
- Correlate VDR expression with West African ancestry
- Examine CYP27B1, CYP24A1, GC expression patterns
- Integrate methylation at VDR locus with expression
- Control for drug treatment effects in analysis

---

5. GWAS Catalog Studies - Vitamin D and Type 2 Diabetes in African Ancestry

Repository: NHGRI-EBI GWAS Catalog
Direct URL: https://www.ebi.ac.uk/gwas/

Key Studies Identified:

Study A: Cross-Ancestry GWAS for 25-Hydroxyvitamin D (25OHD)

Publication: PLoS Genetics (based on search results)
Sample Size: 442,435 UK Biobank participants
- **African ancestry: 8,306 individuals**
- European ancestry: majority of cohort

Key Findings:
- Novel African-specific variant: **rs146759773**
- Low MAF in Europeans, significant in Africans
- GC gene variants (rs7041, rs4588) replicated
- Genome-wide significant dominance effects
- Interactions with skin color

Data Type:
- Summary statistics (beta coefficients, p-values, effect sizes)
- SNP-level associations with 25OHD levels

Access:
- **Public** - Download summary statistics from GWAS Catalog
- Search for publication DOI or lead author
- Filter by trait: “vitamin D levels” or “25-hydroxyvitamin D”

Download URL: https://www.ebi.ac.uk/gwas/downloads/summary-statistics

Study B: Vitamin D Binding Protein (VDBP) GWAS in African Ancestry

Sample Size:
- Southern Community Cohort Study: 2,602 African ancestry
- UK Biobank: 6,934 African/Caribbean ancestry
- **Total: ~9,536 samples**

Key Findings:
- Four GC loci associated with VDBP: rs7041, rs842998, rs8427873, rs11731496
- rs7041 remains significant after conditional analysis

- rs4588 associated with 25OHD concentration
- Effect sizes consistent across African ancestry groups

Data Type:
- GWAS summary statistics
- Protein QTL data

Access:
- Search GWAS Catalog for publication
- May also be in supplementary materials of original paper

Study C: Skin Pigmentation and Vitamin D Deficiency in African Americans

Sample Size: 1,076 African Americans (discovery + replication)

Key Findings:
- **SLC24A5** variants (rs2675345): strongest association ($P=4.0\times10^{-30}$)
- SLC45A2, OCA2 also associated with melanin index
- 11% variance in skin pigmentation explained
- West African ancestry: 23% variance contribution
- Genetic score associated with vitamin D deficiency (OR=1.30)

Data Type:
- GWAS summary statistics
- Melanin index measurements
- Vitamin D levels
- Admixture data

Access:
- PLoS Genetics publication
- Supplementary data with variant details

Study D: Type 2 Diabetes GWAS in African Populations

Study: Meta-analysis including African populations
Sample Size: Thousands across multiple African cohorts

Key Findings:
- **TCF7L2** locus strongest signal (rs7903146, $p=5.3\times10^{-13}$)
- Novel African-specific variant near **AGMO** (rs73284431, $p=5.2\times10^{-9}$)
- 21 loci with shared causal variants across ancestries
- 6 of 100 European-identified loci replicate in Africans

Data Type:
- Meta-analysis summary statistics
- Trans-ethnic fine-mapping results

Access:
- Published journal supplementary materials
- Contact authors for full summary statistics
- May be in dbGaP as summary-level data

General GWAS Catalog Download Instructions:
1. Visit https://www.ebi.ac.uk/gwas/
2. Search by:

- Trait: “vitamin D”, “25-hydroxyvitamin D”, “type 2 diabetes”
- Population: Filter for African ancestry
3. Download options:
- Full summary statistics (when available)
- Top associations table
- Manhattan/QQ plots
4. File formats: Tab-delimited text, JSON
5. Programmatic access via REST API:
`https://www.ebi.ac.uk/gwas/rest/api/studies`

---

6. Additional Genomics Resources

UK Biobank - African/African-Caribbean Subset

URL: https://www.ukbiobank.ac.uk/

Population:
- ~8,000-10,000 individuals of African/African-Caribbean ancestry
- Part of 500,000+ participant cohort

Data Available:
- Genotyping array data
- Whole exome sequencing (50,000 release, expanding)
- Phenotypes: Vitamin D levels, diabetes status, BMI, etc.

Access:
- Requires UK Biobank application
- Research proposal and fee required
- Processing time: 2-3 months

Relevance:
- Vitamin D levels measured (serum 25OHD)
- Type 2 diabetes diagnoses (ICD codes)
- VDR, GC polymorphisms can be extracted
- Rich phenotypic data for covariates

NHLBI CARE Studies (Candidate Gene Association Resource)

URL: Available through dbGaP

Key Cohorts with African American Data:
- Atherosclerosis Risk in Communities (ARIC) Study
- Jackson Heart Study (JHS)
- Multi-Ethnic Study of Atherosclerosis (MESA)

Sample Size: >9,000 African Americans combined

Data Type:
- Candidate gene sequencing
- GWAS data
- Cardiovascular and metabolic phenotypes

Relevance:
- Type 2 diabetes outcomes

- Polygenic risk scores constructed for African Americans
- Gene-diet-physical activity interactions

Access: Controlled access via dbGaP

PROTEOMICS DATA (Priority 2)

1. Metabolic Syndrome Proteomics - MetS Risk Prediction

Dataset ID: PXD039236, PXD039231, PXD038253
Repository: ProteomeXchange Consortium / PRIDE Archive
Direct URL: https://www.ebi.ac.uk/pride/archive/projects/PXD039236

Data Type:
- Serum proteomics
- Data-independent acquisition mass spectrometry (DIA-MS)
- Quantitative proteomics (400+ proteins)

Platforms:
- Mass spectrometry-based proteomics
- DIA-MS methodology

Population:
- Longitudinal cohort
- **Sample size: Nearly 20,000 samples**
- Population not explicitly stated as African ancestry, but metabolic syndrome is relevant

Sample Characteristics:
- Serum samples from participants at risk for metabolic syndrome
- Cases and controls
- Longitudinal follow-up data

Relevance to Research:
- ✓ Insulin signaling proteins
- ✓ Inflammatory markers
- ✓ Apolipoproteins and lipid metabolism
- ✓ Coagulation factors
- Metabolic syndrome (overlap with T2D)
- Machine learning-based risk prediction

Proteins Identified:
- Apolipoproteins (APOA1, APOB, APOE)
- Inflammatory markers
- Coagulation-related factors
- >400 proteins quantified

Data Available:
- Raw MS data files
- Protein identification and quantification tables
- Metadata (sample characteristics, clinical data)
- Machine learning model features

Access Method:
- **Public access** - No application required
- Direct download from PRIDE

File Formats:
- RAW (Thermo Fisher raw files)
- mzML (open format for MS data)
- Pride XML
- Result files (tab-delimited protein quantification)

Preprocessing Information:
- DIA-MS data processing
- Protein inference and quantification
- Normalization applied
- Quality control filters

Download Instructions:
1. Visit PRIDE Archive: https://www.ebi.ac.uk/pride/archive/
2. Search for project accession: PXD039236
3. Download options:
- Individual files (RAW, mzML)
- Complete project via FTP
- Use PRIDE API for programmatic access
4. Tools for analysis:
- MaxQuant, Proteome Discoverer, or Skyline for processing
- R packages: MSstats, DEP for differential analysis
5. File sizes: Typically 100-500 MB per sample (RAW files)

Additional Datasets: PXD039231, PXD038253 (companion studies)

---

2. PRIDE Archive - General Proteomics Repository

Repository: PRIDE (PRoteomics IDEntifications Database)
Direct URL: https://www.ebi.ac.uk/pride/

Description:
- Major public repository for proteomics data
- Part of ProteomeXchange Consortium
- ELIXIR Core Data Resource

Search Strategy for Relevant Datasets:
1. **Search terms:**
- “vitamin D binding protein”
- “type 2 diabetes”
- “insulin signaling”
- “African” or “Black” (population descriptor)
- “serum proteomics” or “plasma proteomics”

1. **Filter by:**
   - Organism: Homo sapiens

- Sample type: Serum, plasma, tissue
- Disease: Diabetes mellitus, metabolic syndrome

2. **Key datasets to explore:**
   - HUPO Plasma Proteome Project data
   - Diabetes-related proteomics studies
   - Inflammatory marker profiling

Data Types Available:
- Mass spectrometry raw files
- Protein identification files
- Peptide-spectrum matches
- Quantification data

Access Method:
- **Public access** for most datasets
- Some may require registration

Tools:
- PRIDE Inspector (desktop tool)
- Web interface for browsing
- FTP access for bulk downloads
- API for programmatic queries

Download Instructions:
1. Browse PRIDE Archive
2. Use advanced search with filters
3. Select projects of interest
4. Download via web interface or FTP
5. Citation required when using data

---

3. Vitamin D Binding Protein (DBP) - Literature-Based Dataset Locations

Based on Search Results:

Study A: Myocardial Infarction Biomarkers

Sample Type: Serum proteomics
Methods: Isotope-coded affinity tags, tandem MS
Findings: Elevated DBP in myocardial infarction

Potential Data Location:
- Check supplementary materials of original publications
- May be in ProteomeXchange under cardiovascular disease

Study B: Bone Mineral Density Study

Sample Type: Serum
Population: Postmenopausal women
Methods: Proteomic profiling
Findings: Increased DBP expression in low bone density

Relevance: Links vitamin D system to bone health and T2D

4. Inflammatory Markers in African Ancestry - Proteomic Studies

Based on Search Results:

Study A: Prostate Cancer Serum Proteomics

Populations:
- African Americans
- Ghanaians
- European Americans

Sample Size: Multiple cohorts (exact numbers in publications)

Key Findings:
- Elevated immune suppression markers in African ancestry
- Chemotaxis proteins (IL-8, CCL23) higher in Africans
- Pleiotrophin, TNFRSF9 upregulated
- Correlation with disease outcomes

Data Type:
- Targeted proteomics (selected reaction monitoring)
- Tandem mass tag-based MS

Access:
- Original publication supplementary materials
- May be deposited in ProteomeXchange/PRIDE
- Contact authors if not publicly available

Study B: Alzheimer’s Disease CSF Proteomics

Populations:
- African Americans
- Caucasians

Sample Type: Cerebrospinal fluid (CSF)

Key Findings:
- Tau and amyloid-beta alterations differ by ancestry
- Inflammatory marker profiles distinct
- Synaptic proteins (VGF, SCG2, NPTX2) vary

Data Type:
- Quantitative proteomics
- CSF protein profiling

Access:
- Check supplementary data in publications
- Possibly in ProteomeXchange

Study C: HIV and Aging Inflammatory Markers

Population: South African cohort
Sample Type: Plasma/serum

Key Findings:
- Inflammation-related proteins (CST5, CCL23)

- Association with African ancestry
- Age-related disease vulnerability

---

5. Insulin Signaling Proteomics - T2D Relevant Studies

Study Type: Phosphoproteomics of insulin resistance

Key Datasets to Search:
- ProteomeXchange for “insulin signaling”
- “phosphoproteomics” + “diabetes”
- “insulin resistance” + “proteomics”

Data Types:
- Phosphoproteomics (PTM profiling)
- Global proteome changes
- Temporal signaling dynamics

Relevant Pathways:
- PI3K-AKT signaling
- ERK pathway
- GSK3 regulation
- Insulin receptor substrate proteins

Recommended Search:
1. PRIDE Archive: Search “insulin resistance” + “Homo sapiens”
2. Filter for serum/plasma samples
3. Look for multi-ethnic or African ancestry studies

METABOLOMICS DATA (Priority 3)

1. Nigerian Type 2 Diabetes Metabolomics (AADM Study)

Dataset ID: Associated with AADM genomic study (phs001844)
Repository: Published data (check supplementary materials)
Publication: Genome Medicine, 2024

Data Type:
- Untargeted metabolomics
- Plasma metabolite profiling
- >1,000 metabolites analyzed

Platform:
- Mass spectrometry-based metabolomics
- Likely LC-MS/MS

Population:
- **Nigerian participants**
- Type 2 diabetes cases and controls
- Part of Africa America Diabetes Mellitus (AADM) study

Sample Size:
- Large cohort from Sub-Saharan Africa
- Discovery and replication cohorts

Metabolite Categories:
- 280 differentially expressed metabolites (DEMs)
- 51% lipid pathways
- 21% amino acids
- Carbohydrates
- Bile acids

Relevance to Research:
- ✓ Glucose metabolism pathways (glycolysis)
- ✓ Free fatty acid metabolism
- ✓ Branched-chain amino acid catabolism
- ✓ Bile acid metabolism
- Type 2 diabetes biomarkers
- African ancestry-specific metabolic profiles

Key Findings:
- 10-metabolite biomarker panel for T2D prediction
- Includes: glucose, gluconate, mannose, metformin, 1,5-anhydroglucitol
- AUC 0.924 (discovery), 0.935 (replication)
- Correlations with HbA1c, insulin resistance

Data Available:
- Metabolite abundance tables
- Differential expression results
- Biomarker panel composition
- Clinical correlations

Access Method:
- Check publication supplementary materials: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-024-01308-5
- Data may be deposited in Metabolomics Workbench
- Contact authors if not publicly posted

File Formats:
- Excel/CSV (metabolite tables)
- Possibly mzML (raw MS data)

Download Instructions:
1. Access publication: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-024-01308-5
2. Download supplementary data files
3. Check for data repository links in paper
4. If raw data needed, contact corresponding author

---

2. South African Women T2D Development - Longitudinal Metabolomics

Dataset ID: v8gbcx9gp2.1
Repository: Mendeley Data
DOI: 10.17632/v8gbcx9gp2.1
Direct URL: https://data.mendeley.com/datasets/v8gbcx9gp2/1

Data Type:
- Metabolomics data
- Baseline and follow-up measurements
- Longitudinal study design

Platform:
- Mass spectrometry-based metabolomics

Population:
- **Black South African women**
- Participants developing Type 2 diabetes
- Longitudinal follow-up

Sample Characteristics:
- Baseline samples (pre-diabetes)
- Follow-up samples (T2D development)
- Matched controls

Metabolite Categories:
- Phospholipids (especially lysophospholipids)
- Bile acids
- Branched-chain amino acids (BCAAs)
- Lipid metabolism intermediates

Relevance to Research:
- ✓ Metabolic changes during T2D progression
- ✓ Early biomarkers of diabetes risk
- ✓ Lipid metabolism alterations
- ✓ Amino acid catabolism
- ✓ African ancestry-specific metabolic trajectories

Data Available:
- Two separate data sheets (baseline and follow-up)
- Metabolite abundance values
- Sample metadata
- Clinical measurements

Access Method:
- **Public access** - Open license (CC BY 4.0)
- Direct download from Mendeley Data

File Formats:
- Excel (.xlsx) or CSV
- Separate sheets for baseline and follow-up

Contributor: Elin Chorell
Publication Date: March 31, 2020

Categories: Metabolomics, Type 2 Diabetes, Ethnicity, Insulin Resistance

Download Instructions:
1. Visit: https://data.mendeley.com/datasets/v8gbcx9gp2/1
2. Click “Download” button
3. No registration required (CC BY 4.0 license)
4. Cite DOI when using data: 10.17632/v8gbcx9gp2.1
5. Read associated publication for methodology

---

3. NIH Metabolomics Workbench - General Repository

Repository: National Metabolomics Data Repository (NMDR)
Direct URL: https://www.metabolomicsworkbench.org/

Repository Statistics (as of Sept 30, 2025):
- **4,150 studies total**
- 3,739 publicly available
- 411 embargoed

Data Types:
- LC-MS metabolomics
- GC-MS metabolomics
- NMR spectroscopy
- Lipidomics
- Targeted and untargeted approaches

Search Strategy for Relevant Datasets:

Keywords to search:
1. “vitamin D” or “25-hydroxyvitamin D” or “calcitriol”
2. “diabetes” or “glucose” or “insulin resistance”
3. “African” or “Black” or “African American”
4. Combinations: “vitamin D” AND “diabetes”

Filtering Options:
- Species: Homo sapiens
- Sample source: Blood, serum, plasma
- Analysis type: LC-MS, GC-MS
- Study factors: Disease state, ethnicity

Specific Study Identified:

Study ID: ST002681

Title: T2D prediction metabolomics
Description: Metabolomics for type 2 diabetes risk prediction
Sample Type: Human serum/plasma
Methods: LC-MS

Relevance:
- Type 2 diabetes biomarkers
- Machine learning prediction
- Metabolic risk profiling

Access:
- Public study in NMDR
- Search by study ID: ST002681
- Direct URL: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&StudyID=ST002681

---

4. Vitamin D Metabolites Studies in Workbench

Search Results: Multiple studies on vitamin D metabolomes

Study Type A: Vitamin D and Lipid Metabolism

Sample Type: Serum
Key Metabolites:
- 25(OH)D (vitamin D status marker)
- CMPF (3-carboxy-4-methyl-5-propyl-2-furanpropanoic acid)
- EPA, DHA (omega-3 fatty acids)
- Phospholipids (GPPE)

Associations:
- Positive correlation with lipid metabolites
- Inverse correlation with certain phospholipids
- Links to cardiovascular health

Population: Various cohorts (check individual studies)

Study Type B: Vitamin D Insufficiency Metabolomics

Key Findings:
- Altered lipid profiles
- Acylcarnitine changes
- Amino acid metabolism
- Associations with dyslipidemia

Relevance:
- Vitamin D deficiency metabolic consequences
- May overlap with T2D metabolic signatures

---

5. Glucose Metabolism in African Ancestry - Search Results

Studies Identified from Search:

Study A: African American vs Caucasian Metabolomics

Repository: May be in Metabolomics Workbench
Population: African Americans, comparisons with other groups

Key Findings:
- Elevated fasting glucose associated with African ancestry
- Higher HbA1c levels
- Genetic × metabolic interactions
- Admixture effects on glucose regulation

Metabolite Classes:
- Glucose and glucose derivatives
- Amino acids (especially BCAAs)
- Lipids (glycerophospholipids, sphingolipids)
- TCA cycle intermediates

Study B: Hypertension Metabolomics in West African Ancestry

Key Metabolite Clusters:
- Plasmalogen/lysoplasmalogen
- Sphingolipid metabolism
- Urea cycle metabolites
- Associations with genetic ancestry

Relevance:
- Overlap between hypertension and T2D
- African ancestry-specific metabolic profiles
- Cardiovascular-metabolic connections

---

6. Additional Metabolomics Resources

Human Metabolome Database (HMDB)

URL: https://www.hmdb.ca/

Description:
- Comprehensive metabolite reference database
- Links to pathways and proteins
- Metabolite IDs and structures

Use for Research:
- Annotate metabolites from studies
- Link metabolites to vitamin D and glucose pathways
- Identify pathway connections

Access: Public, free

MetaboLights

URL: https://www.ebi.ac.uk/metabolights/

Description:
- Open-access metabolomics repository
- EMBL-EBI hosted
- >4,000 studies

Search Strategy:
- Search for diabetes, vitamin D, African populations

- Filter by organism and sample type
- Download raw data when available

Access: Public

Study Search Tips:

1. Use terms: “type 2 diabetes”, “glucose metabolism”, “insulin resistance”
2. Filter for human studies
3. Look for plasma/serum samples
4. Check for ethnicity metadata
5. Download raw data (mzML, mzXML) when available

QUICK REFERENCE SUMMARY

Highest Priority Datasets for Immediate Download

Dataset	Type	ID/Accession	Population	Access	Priority
AADM GWAS	Genomics	phs001844	Sub-Saharan African	Controlled	★★★★★
T2D-GENES	Genomics	phs001610	Multi-ethnic (6,468 African Americans)	Controlled	★★★★★
GSE124076	Genomics (RNA-seq)	GSE124076	African American	Public	★★★★★
South African T2D	Metabolomics	v8gbcx9gp2	Black South African	Public	★★★★★
Nigerian AADM	Metabolomics	In publication	Nigerian	Public (supp)	★★★★
MetS Proteomics	Proteomics	PXD039236	General cohort	Public	★★★

Data by Omics Layer

GENOMICS (6 key datasets)
- 3 dbGaP studies (controlled access)
- 1 GEO study (public)
- GWAS Catalog studies (public)
- UK Biobank (application required)

PROTEOMICS (5+ datasets)
- 3 ProteomeXchange studies (public)

- PRIDE repository (public, search required)
- Study-specific data (check publications)

METABOLOMICS (6+ datasets)
- 2 confirmed public datasets
- Metabolomics Workbench (search required)
- MetaboLights (search required)
- Publication supplementary data

DATA INTEGRATION STRATEGY

Phase 1: Immediate Downloads (Week 1-2)

1. **Public datasets (no barriers):**
   - GSE124076 (gene expression)
   - Mendeley South African metabolomics
   - GWAS Catalog summary statistics
   - PRIDE proteomics (PXD039236, etc.)

2. **Start controlled access applications:**
   - dbGaP account creation
   - DAR preparation for phs001844
   - DAR preparation for phs001610
   - IRB documentation

Phase 2: Controlled Access Datasets (Week 3-8)

1. **Process dbGaP applications:**
   - Submit DARs with IRB approval
   - Await approval (2-4 weeks typical)
   - Set up download infrastructure

2. **Download large-scale data:**
   - AADM genotypes and phenotypes
   - T2D-GENES exome sequences
   - African American subsets

Phase 3: Data Processing and Integration (Week 9+)

1. **Genomics layer:**
   - Extract VDR, GC, CYP27B1, CYP24A1 variants
   - Identify T2D susceptibility loci
   - Calculate polygenic risk scores
   - Admixture analysis

2. **Gene expression layer:**
   - Quantify VDR, GC, CYP gene expression
   - Correlate with ancestry proportions
   - Identify eQTLs for vitamin D genes
   - Pathway enrichment analysis

3. **Proteomics layer:**
   - Quantify vitamin D binding protein
   - Insulin signaling proteins
   - Inflammatory markers
   - Link to genomic variants

4. **Metabolomics layer:**
   - Vitamin D metabolites (25OHD, 1,25(OH)2D)
   - Glucose metabolism intermediates
   - Lipid profiles
   - Amino acid profiles

Hierarchical Integration Plan

```
LEVEL 1: GENOMICS
├── Vitamin D pathway SNPs (VDR, GC, CYP27B1, CYP24A1)
├── T2D susceptibility loci (TCF7L2, etc.)
└── African ancestry markers
         ↓
LEVEL 2: TRANSCRIPTOMICS
├── VDR gene expression
├── Vitamin D metabolism genes
└── Insulin signaling genes
         ↓
LEVEL 3: PROTEOMICS
├── Vitamin D binding protein (DBP)
├── Insulin signaling proteins
└── Inflammatory markers
         ↓
LEVEL 4: METABOLOMICS
├── 25(OH)D levels
├── Glucose and insulin
├── Lipid profiles
└── BCAAs and amino acids
         ↓
INTEGRATED ANALYSIS
├── SNP → gene expression associations
├── Expression → protein abundance
├── Protein → metabolite levels
└── Multi-omics → T2D phenotype
```

DOWNLOAD CHECKLIST

Immediate Actions

- [ ] Create GEO account
- [ ] Download GSE124076 RNA-seq data
- [ ] Download Mendeley metabolomics dataset (v8gbcx9gp2)
- [ ] Search GWAS Catalog for vitamin D studies
- [ ] Download GWAS summary statistics
- [ ] Search PRIDE for vitamin D binding protein studies
- [ ] Download PXD039236 proteomics data

Within 1 Week

- [ ] Create dbGaP account
- [ ] Prepare IRB documentation
- [ ] Draft Data Access Request for phs001844
- [ ] Draft Data Access Request for phs001610
- [ ] Identify institutional signing official
- [ ] Search Metabolomics Workbench for relevant studies
- [ ] Create list of secondary datasets

Within 1 Month

- [ ] Submit dbGaP DARs
- [ ] Track approval status
- [ ] Set up compute environment for large files
- [ ] Download additional GEO datasets if identified
- [ ] Search for additional proteomics studies
- [ ] Document all data sources in lab notebook

Post-Approval (2-3 Months)

- [ ] Download AADM GWAS data (phs001844)
- [ ] Download T2D-GENES exome data (phs001610)
- [ ] Verify data integrity
- [ ] Begin QC and preprocessing
- [ ] Document provenance of all datasets

COMPUTATIONAL REQUIREMENTS

Storage Needs

- **Genomics:**
- GWAS: ~10-50 GB (genotypes + phenotypes)
- WES: ~50-100 GB per sample × thousands = TB-scale
- RNA-seq: ~5-10 GB per sample
- **Proteomics:**
- Raw MS: ~100-500 MB per sample
- Processed: <100 MB
- **Metabolomics:**
- Generally smaller, <10 GB total

Total estimated: 2-5 TB (depending on subsets)

Software Requirements

- **Genomics:**
- PLINK (GWAS analysis)
- GATK (variant calling)
- bcftools, vcftools
- ADMIXTURE (ancestry)

- R/Bioconductor packages
- **Transcriptomics:**
- GEOquery (R package)
- DESeq2, edgeR
- STAR or Salmon (alignment/quantification)
- **Proteomics:**
- MaxQuant or Proteome Discoverer
- MSstats (R package)
- PRIDE Inspector
- **Metabolomics:**
- XCMS (R package)
- MetaboAnalyst
- MZmine
- **Integration:**
- MultiAssayExperiment (R/Bioconductor)
- Python (pandas, scikit-learn)
- Custom scripts

CITATION REQUIREMENTS

When Using These Datasets:

1. **Cite original publications**
2. **Acknowledge data repositories:**
   - dbGaP studies: Acknowledge NCBI and funding sources
   - GEO: Cite GEO accession and original submitters
   - GWAS Catalog: Acknowledge NHGRI-EBI
   - ProteomeXchange: Cite PRIDE and dataset IDs
   - Metabolomics Workbench: Acknowledge NIH Common Fund
3. **Example acknowledgment:**

> "This study used data from the Africa America Diabetes Mellitus (AADM) study (phs001844), obtained through dbGaP, the T2D-GENES consortium (phs001610), gene expression data from GSE124076 available through NCBI GEO, and metabolomics data from Mendeley Data (DOI: 10.17632/v8gbcx9gp2.1) and the NIH Metabolomics Workbench. We thank the participants and investigators of these studies."

CONTACT INFORMATION FOR DATA ACCESS ISSUES

dbGaP Support

- Email: dbgap-help@ncbi.nlm.nih.gov
- Phone: +1-301-451-5245

GEO Support

- Email: geo@ncbi.nlm.nih.gov
- Web: https://www.ncbi.nlm.nih.gov/geo/info/contact.html

PRIDE Support

- Email: pride-support@ebi.ac.uk
- Web: https://www.ebi.ac.uk/pride/markdownpage/contactpage

Metabolomics Workbench

- Email: metabolomics.workbench@gmail.com
- Help: https://www.metabolomicsworkbench.org/about/howtocite.php

---

NOTES AND CAVEATS

1. **Data Access Timelines:**
   - Public data: Immediate
   - Controlled access (dbGaP): 2-4 weeks after complete application
   - Some datasets may require additional institutional agreements

2. **Data Use Limitations:**
   - dbGaP data has specific use restrictions (General Research Use vs. Disease-Specific)
   - Must comply with data use agreements
   - Cannot attempt re-identification of participants
   - Acknowledge in publications

3. **Sample Overlap:**
   - Some participants may be in multiple studies
   - Check for overlap between AADM and T2D-GENES
   - Avoid double-counting in meta-analyses

4. **Population Stratification:**
   - African ancestry is diverse (West African, East African, admixed African American)
   - Account for population structure in all analyses
   - Use appropriate reference panels

5. **Data Quality:**
   - Some older datasets may have lower quality
   - Check sequencing depth, genotyping call rates
   - Verify metabolite identification confidence

6. **Missing Data:**
   - Not all samples have all omics layers

- Some studies lack vitamin D measurements
- May need to impute or subset analyses

7. **Batch Effects:**
   - Data from different centers/platforms may have batch effects
   - Requires careful normalization and adjustment
   - Consider meta-analysis approaches

Document Version: 1.0
Prepared by: AI Research Agent
Date: September 30, 2025
Next Review: Upon completion of data downloads and initial QC

APPENDIX: SEARCH QUERIES USED

Genomics Searches

1. “African ancestry vitamin D GWAS Catalog”
2. “African ancestry type 2 diabetes dbGaP datasets”
3. “VDR gene expression African ancestry GEO”
4. “vitamin D receptor polymorphism African population NCBI”
5. “GC gene CYP27B1 African ancestry genomics data”

Proteomics Searches

1. “vitamin D binding protein proteomics PRIDE”
2. “type 2 diabetes insulin signaling proteomics ProteomeXchange African”
3. “inflammatory markers proteomics African ancestry”
4. “metabolic syndrome proteomics data repository”

Metabolomics Searches

1. “vitamin D metabolites Metabolomics Workbench”
2. “glucose metabolism metabolomics African ancestry”
3. “type 2 diabetes metabolomics data repository African”

END OF INVENTORY