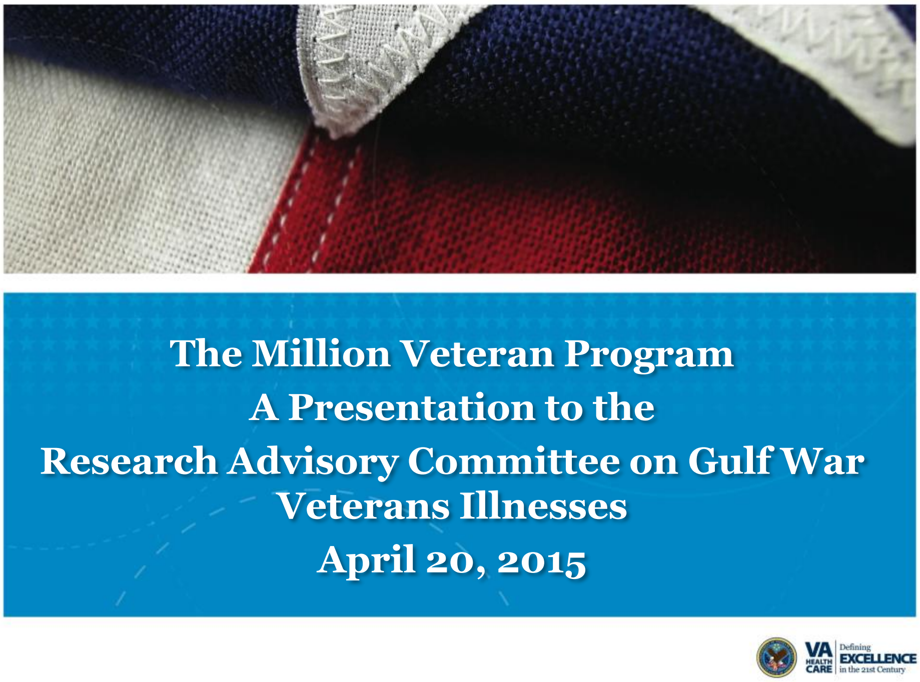
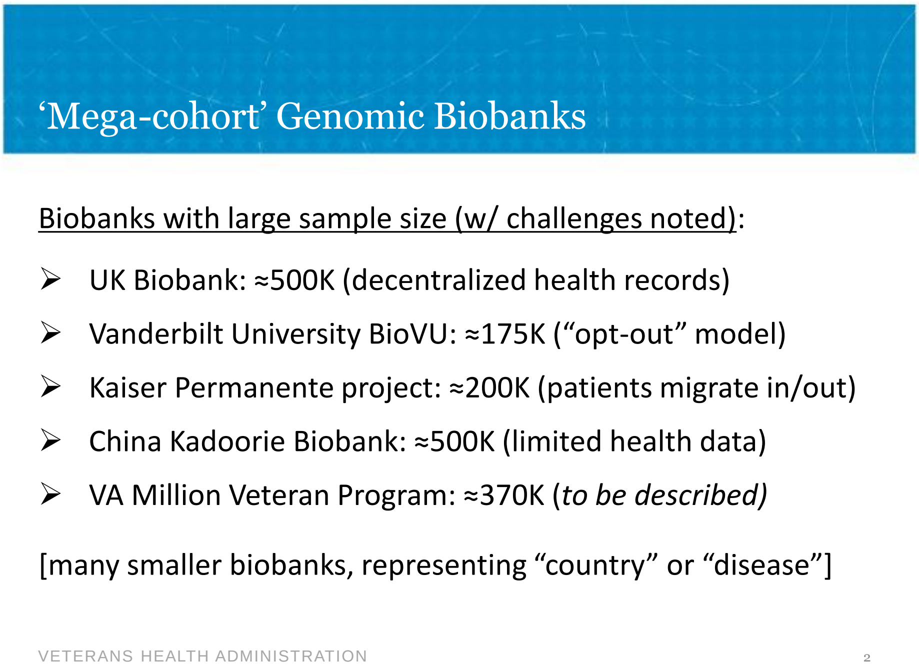


The Million Veteran Program
A Presentation to the
Research Advisory Committee on Gulf War Veterans Illnesses
April 20, 2015

VA HEALTH CARE
Defining EXCELLENCE in the 21st Century

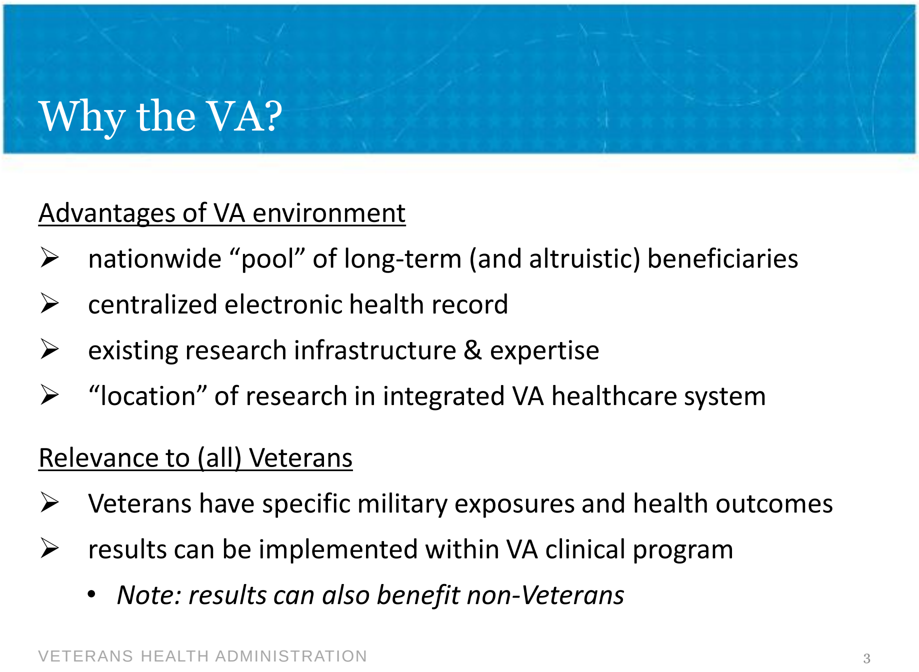


‘Mega-cohort’ Genomic Biobanks

Biobanks with large sample size (w/ challenges noted):

- UK Biobank: ≈500K (decentralized health records)
- Vanderbilt University BioVU: ≈175K (“opt-out” model)
- Kaiser Permanente project: ≈200K (patients migrate in/out)
- China Kadoorie Biobank: ≈500K (limited health data)
- VA Million Veteran Program: ≈370K (*to be described)*

[many smaller biobanks, representing “country” or “disease”]



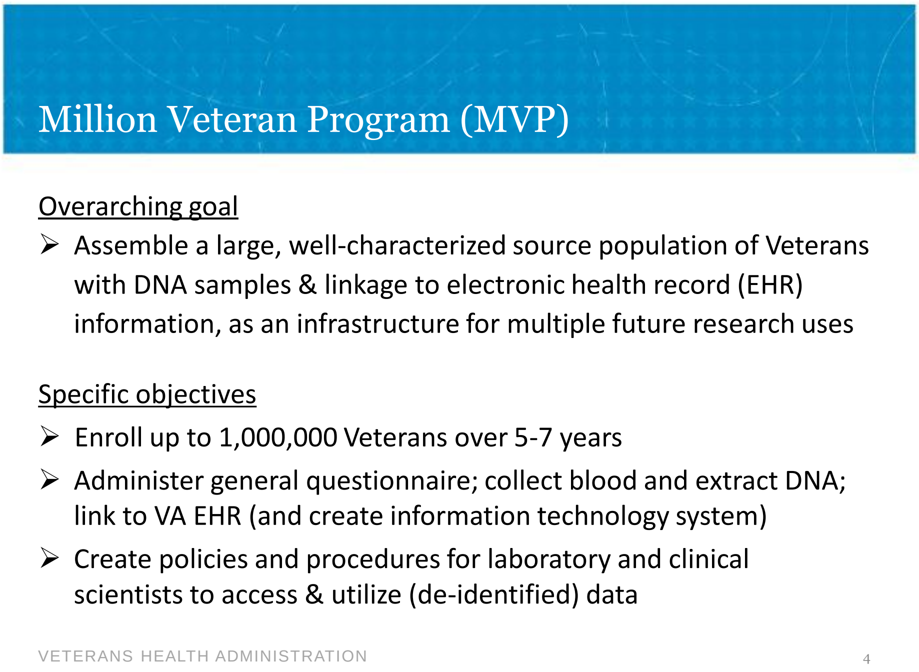
Why the VA?

Advantages of VA environment

- nationwide “pool” of long-term (and altruistic) beneficiaries
- centralized electronic health record
- existing research infrastructure & expertise
- “location” of research in integrated VA healthcare system

Relevance to (all) Veterans

- Veterans have specific military exposures and health outcomes
- results can be implemented within VA clinical program
  - *Note: results can also benefit non-Veterans*



Million Veteran Program (MVP)

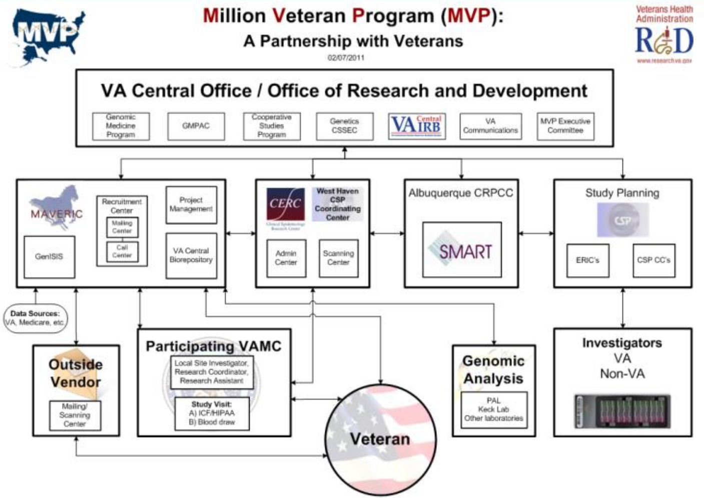
Overarching goal

- Assemble a large, well-characterized source population of Veterans with DNA samples & linkage to electronic health record (EHR) information, as an infrastructure for multiple future research uses

Specific objectives

- Enroll up to 1,000,000 Veterans over 5-7 years
- Administer general questionnaire; collect blood and extract DNA; link to VA EHR (and create information technology system)
- Create policies and procedures for laboratory and clinical scientists to access & utilize (de-identified) data

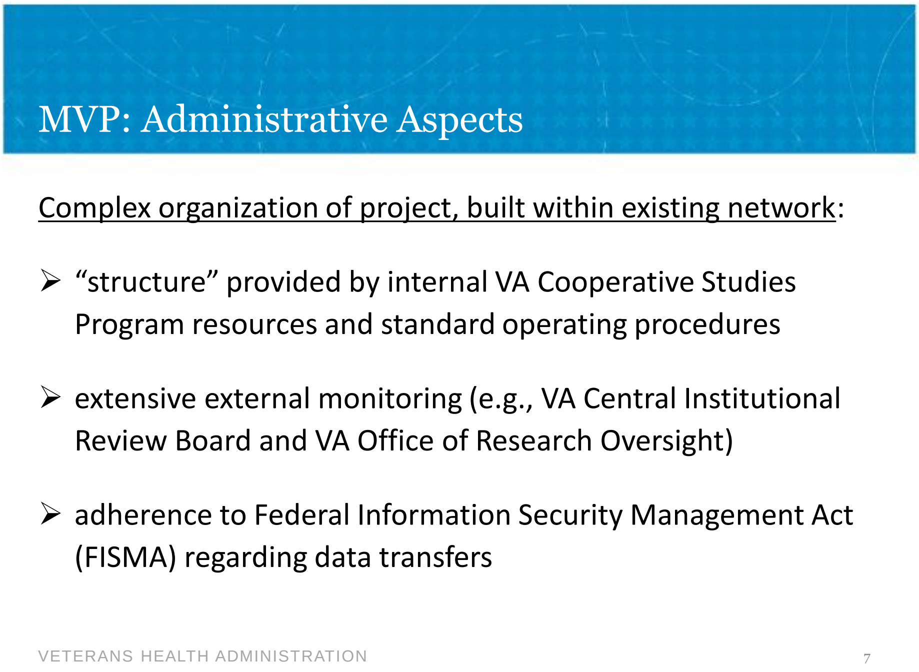
MVP Organizational Structure


Million Veteran Program (MVP):
A Partnership with Veterans
02/07/2011

Scope of MVP

Operational aspects of the Million Veteran Program

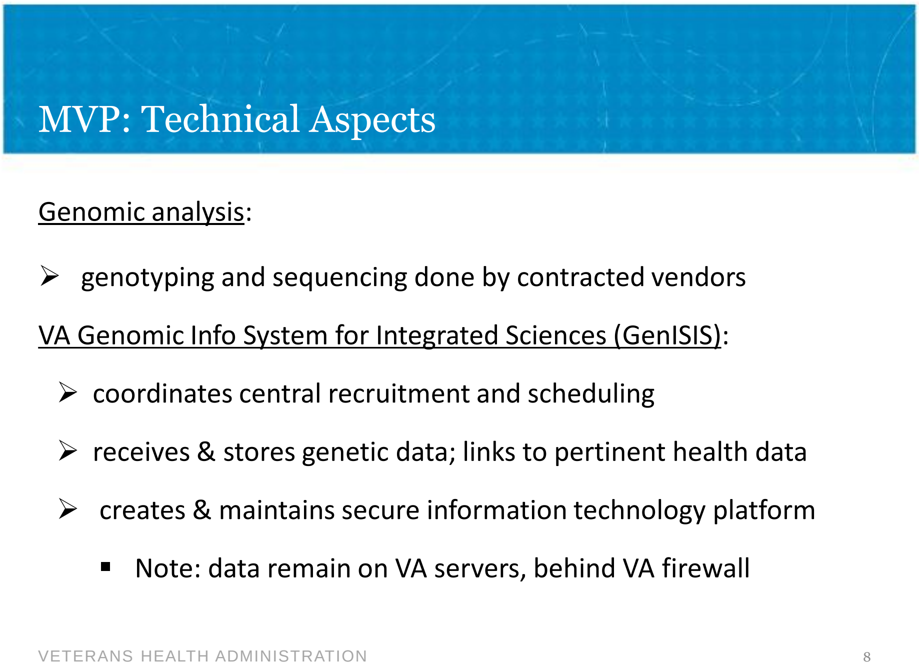
- Recruit using opt-in/decline model
  - invitational letter are being sent to ≈6 million Veteran Health Administration beneficiaries
- Initial enrollees at vanguard sites in 2011 → 50 sites as of 2014
- J. Michael Gaziano & John Concato, Co-Principal Investigators
  - funding from VA Cooperative Studies & Genomic Medicine
- Example of “team science” and “big data” within existing VA infrastructure, involving administrative, technical, ethical, and scientific challenges



MVP: Administrative Aspects

Complex organization of project, built within existing network:

- “structure” provided by internal VA Cooperative Studies Program resources and standard operating procedures
- extensive external monitoring (e.g., VA Central Institutional Review Board and VA Office of Research Oversight)
- adherence to Federal Information Security Management Act (FISMA) regarding data transfers



MVP: Technical Aspects

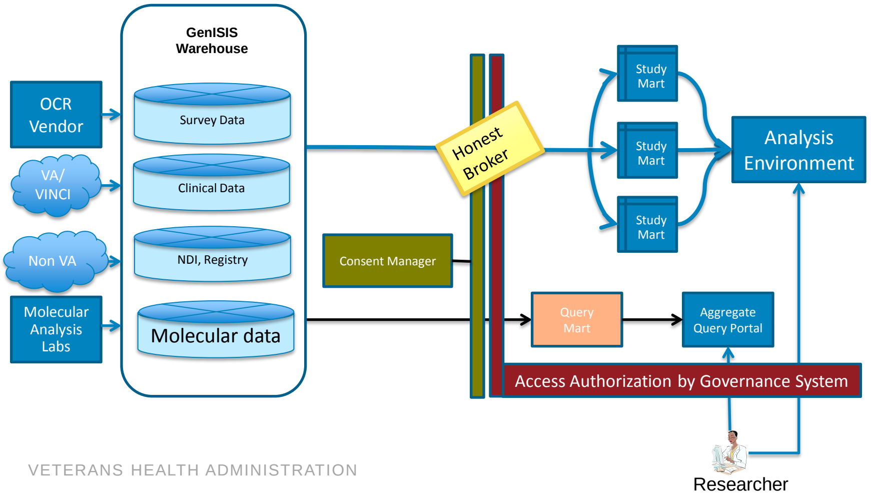
Genomic analysis:

- genotyping and sequencing done by contracted vendors

VA Genomic Info System for Integrated Sciences (GenISIS):

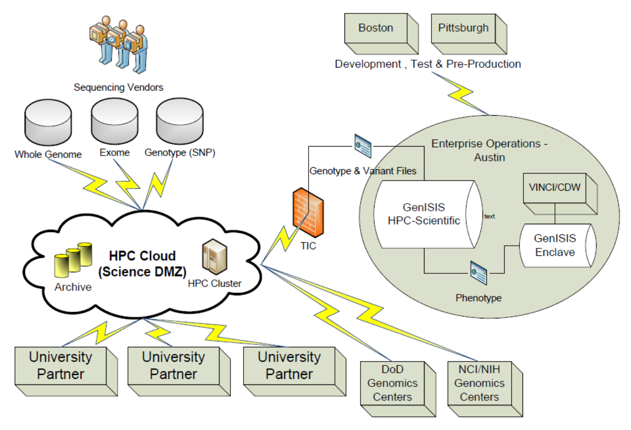
- coordinates central recruitment and scheduling
- receives & stores genetic data; links to pertinent health data
- creates & maintains secure information technology platform
  - Note: data remain on VA servers, behind VA firewall

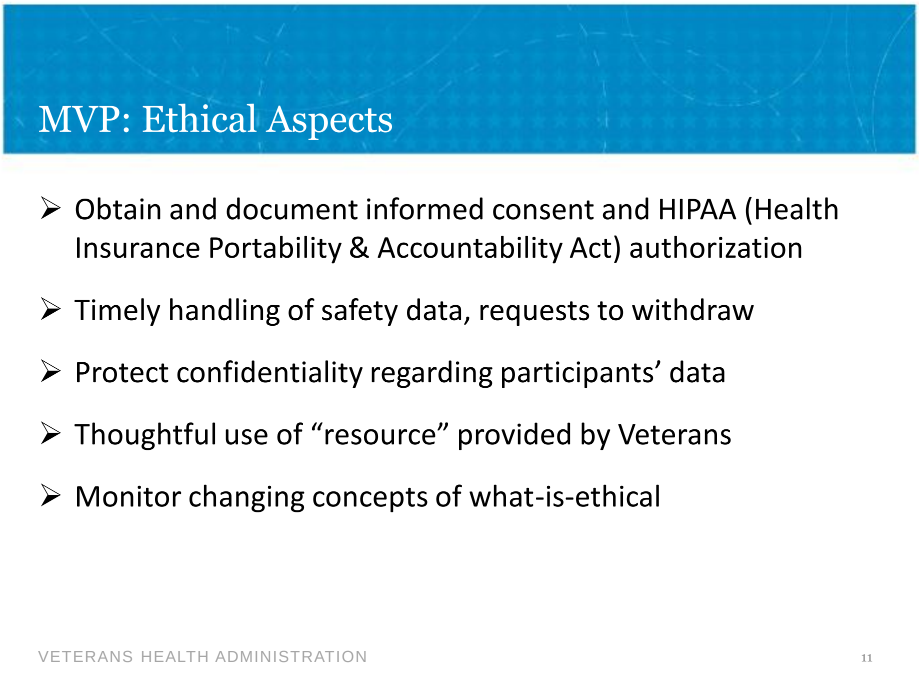
MVP Data Architecture in GenISIS



MVP: Information Technology

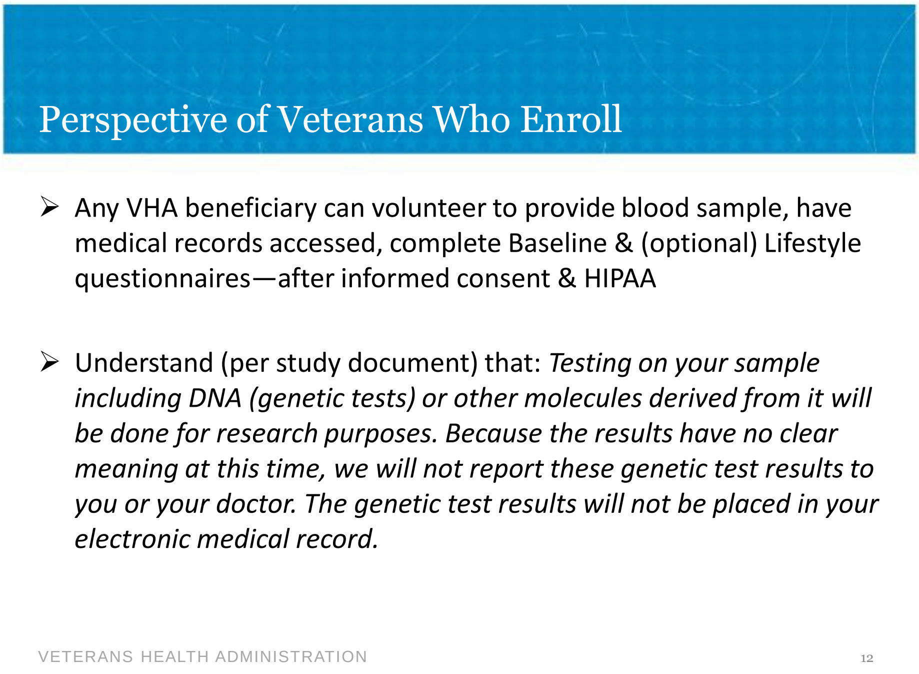
MVP "To-Be" Architecture





MVP: Ethical Aspects

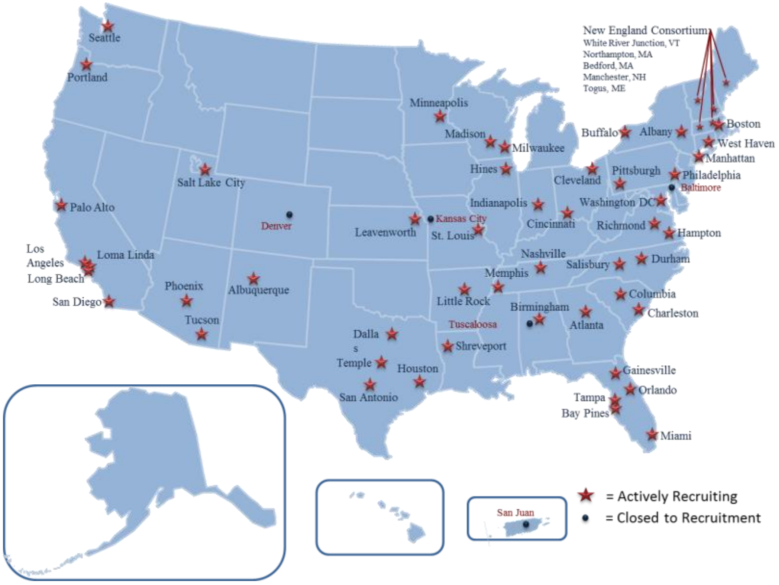
- Obtain and document informed consent and HIPAA (Health Insurance Portability & Accountability Act) authorization
- Timely handling of safety data, requests to withdraw
- Protect confidentiality regarding participants' data
- Thoughtful use of "resource" provided by Veterans
- Monitor changing concepts of what-is-ethical



Perspective of Veterans Who Enroll

- Any VHA beneficiary can volunteer to provide blood sample, have medical records accessed, complete Baseline & (optional) Lifestyle questionnaires—after informed consent & HIPAA
- Understand (per study document) that: *Testing on your sample including DNA (genetic tests) or other molecules derived from it will be done for research purposes. Because the results have no clear meaning at this time, we will not report these genetic test results to you or your doctor. The genetic test results will not be placed in your electronic medical record.*

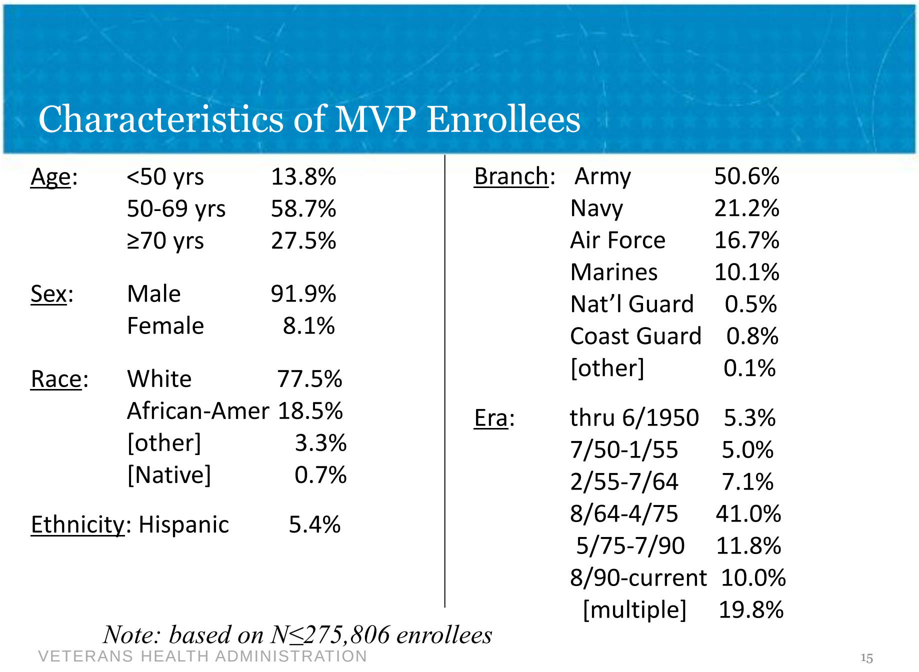
MVP Enrollment Sites



MVP Update

Status as of 27 Feb 2015

invitations mailed	2,758,012
baseline surveys returned	455,150
consent forms (& blood)	**353,380**
specimens sent for (as of end FY14):	
- genotyping	206,303
- exome sequencing	24,260
- whole-genome sequencing	1,886



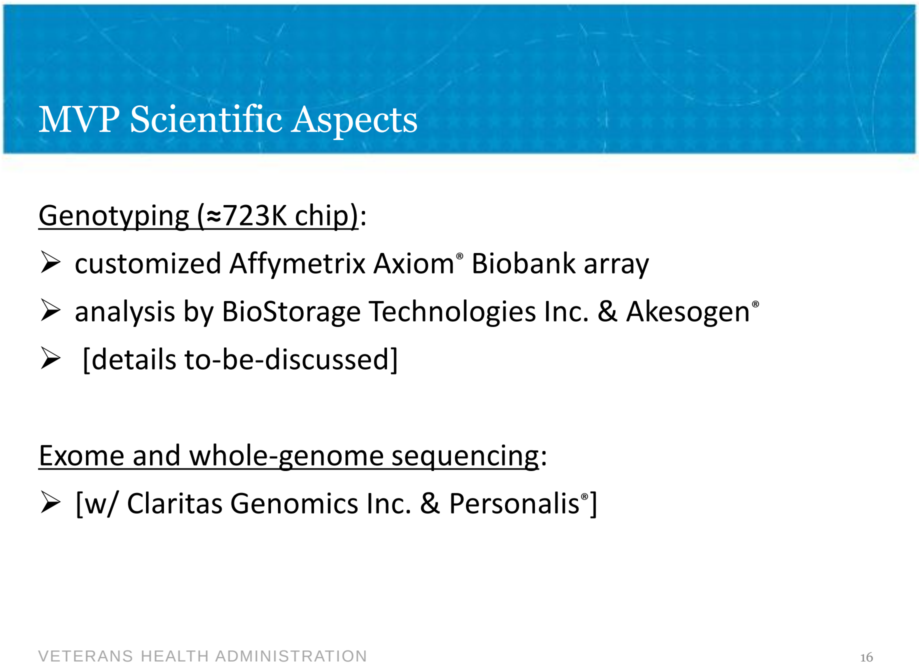
Characteristics of MVP Enrollees

Age:	<50 yrs	13.8%
	50-69 yrs	58.7%
	≥70 yrs	27.5%
Sex:	Male	91.9%
	Female	8.1%
Race:	White	77.5%
	African-Amer	18.5%
	[other]	3.3%
	[Native]	0.7%
Ethnicity:	Hispanic	5.4%

Branch:	Army	50.6%
	Navy	21.2%
	Air Force	16.7%
	Marines	10.1%
	Nat'l Guard	0.5%
	Coast Guard	0.8%
	[other]	0.1%
Era:	thru 6/1950	5.3%
	7/50-1/55	5.0%
	2/55-7/64	7.1%
	8/64-4/75	41.0%
	5/75-7/90	11.8%
	8/90-current	10.0%
	[multiple]	19.8%

Note: based on N≤275,806 enrollees

VETERANS HEALTH ADMINISTRATION 15



MVP Scientific Aspects

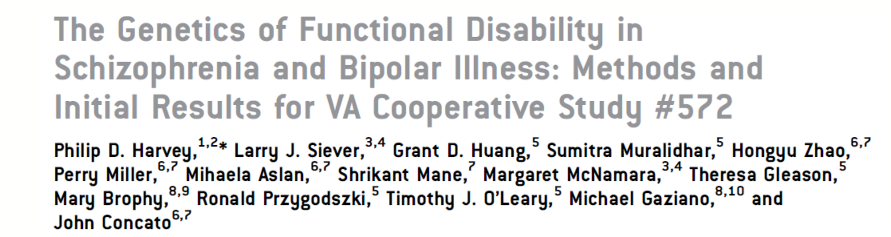
Genotyping (≈723K chip):

- customized Affymetrix Axiom® Biobank array
- analysis by BioStorage Technologies Inc. & Akesogen®
- [details to-be-discussed]

Exome and whole-genome sequencing:

- [w/ Claritas Genomics Inc. & Personalis®]

VETERANS HEALTH ADMINISTRATION 16



The Genetics of Functional Disability in Schizophrenia and Bipolar Illness: Methods and Initial Results for VA Cooperative Study #572

Philip D. Harvey,[1,2]* Larry J. Siever,[3,4] Grant D. Huang,[5] Sumitra Muralidhar,[5] Hongyu Zhao,[6,7] Perry Miller,[6,7] Mihaela Aslan,[6,7] Shrikant Mane,[7] Margaret McNamara,[3,4] Theresa Gleason,[5] Mary Brophy,[8,9] Ronald Przygodszki,[5] Timothy J. O'Leary,[5] Michael Gaziano,[8,10] and John Concato[6,7]

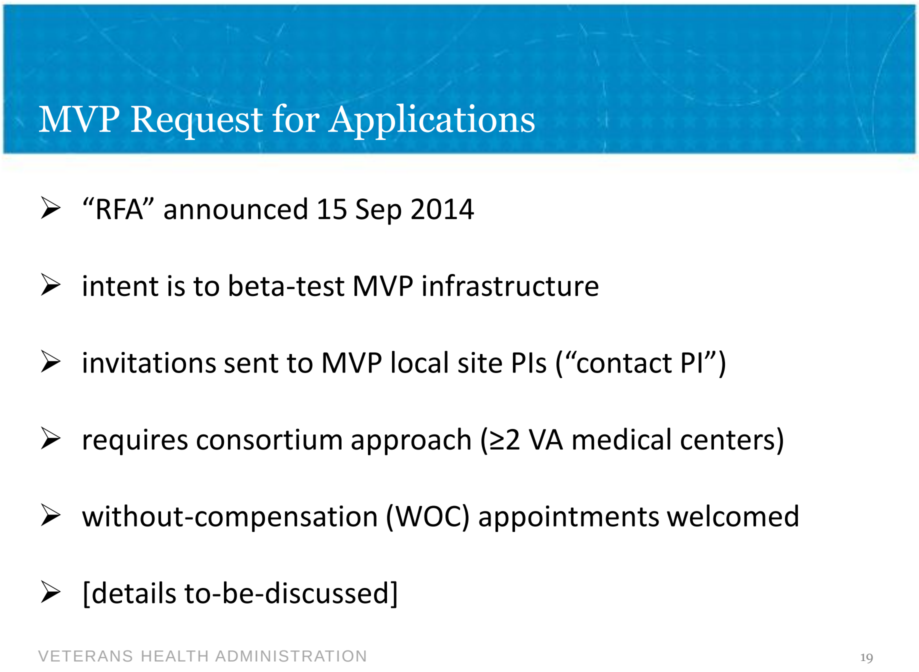
<u>GWAS of schizophrenia and bipolar disorder</u>:

- N=9,355 case patients w/ extensive phenotyping
- control patients to-be-identified from MVP
- genotyping ongoing using MVP chip
- main study is initial alpha-test of entire MVP infrastructure
- sub-study is conducting exome-sequencing (N=600) in collaboration with Yale Center for Genomic Analysis

Second MVP-related Project

<u>Genomics of posttraumatic stress disorder (CSP#575B)</u>:

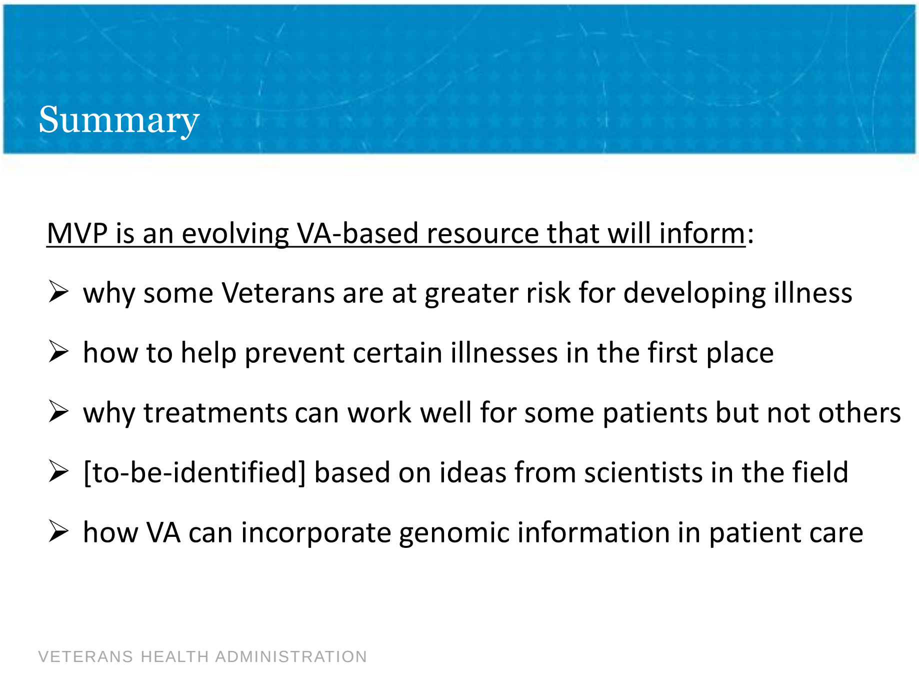
- GWAS of combat-exposed Veterans; funded Dec 2013
- w/ Joel Gelernter (Yale) and Murray Stein (UCSD)
- both case and control patients come from MVP
- phenotyping ongoing w/ electronic records & questionnaire
- genotyping ongoing using MVP chip
- represents α-test project that is completely 'intra-MVP'



MVP Request for Applications

- “RFA” announced 15 Sep 2014
- intent is to beta-test MVP infrastructure
- invitations sent to MVP local site PIs (“contact PI”)
- requires consortium approach (≥2 VA medical centers)
- without-compensation (WOC) appointments welcomed
- [details to-be-discussed]

VETERANS HEALTH ADMINISTRATION 19



Summary

MVP is an evolving VA-based resource that will inform:

- why some Veterans are at greater risk for developing illness
- how to help prevent certain illnesses in the first place
- why treatments can work well for some patients but not others
- [to-be-identified] based on ideas from scientists in the field
- how VA can incorporate genomic information in patient care

VETERANS HEALTH ADMINISTRATION

Genomic Technologies and Contracts

MVP Genomic Analyses

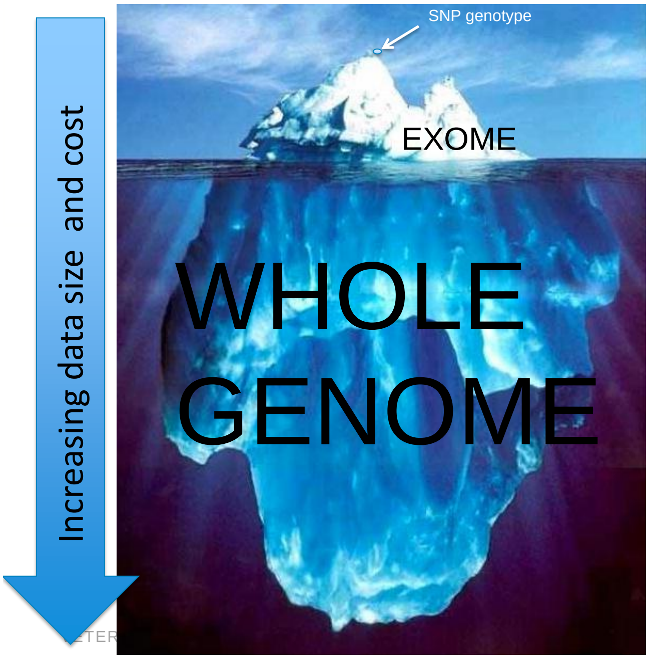
- Genome sequencing
- Exome sequencing
- DNA genotyping
- Overall costs and samples

MVP Genomic Analyses Con’t

Hundreds of thousands of MVP samples coming in the door, but they are only as good as the data we can get out of them

Large scale Centralized contracting

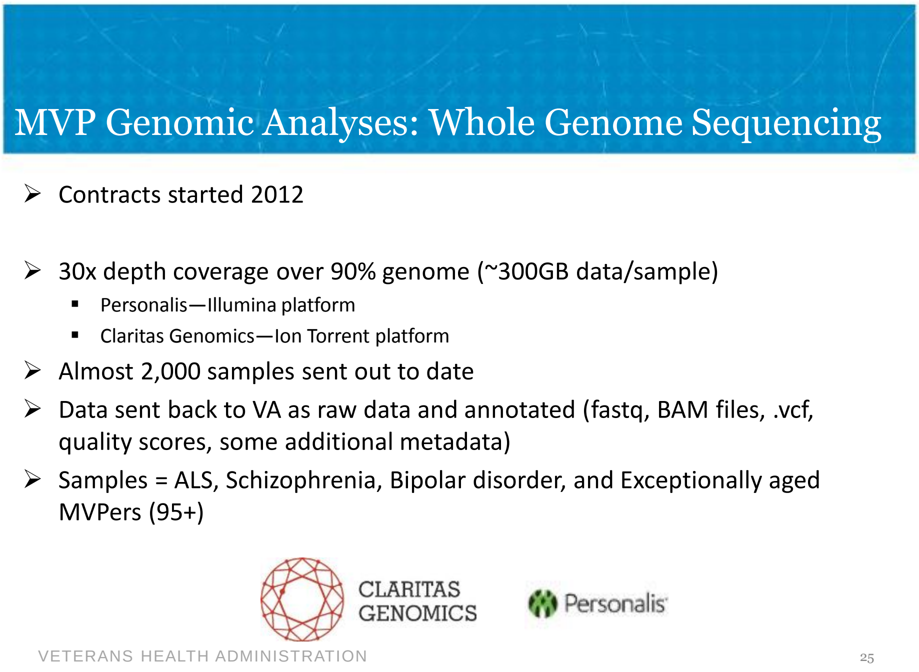
- ➢ Large volumes= better value
- ➢ Standardized (relatively) formats = easier analysis going forward
- ➢ 3 data analysis types



allele differences specific nucleotide locations

DNA sequence in **protein coding** regions of the genome

DNA sequence across the **entire** genome

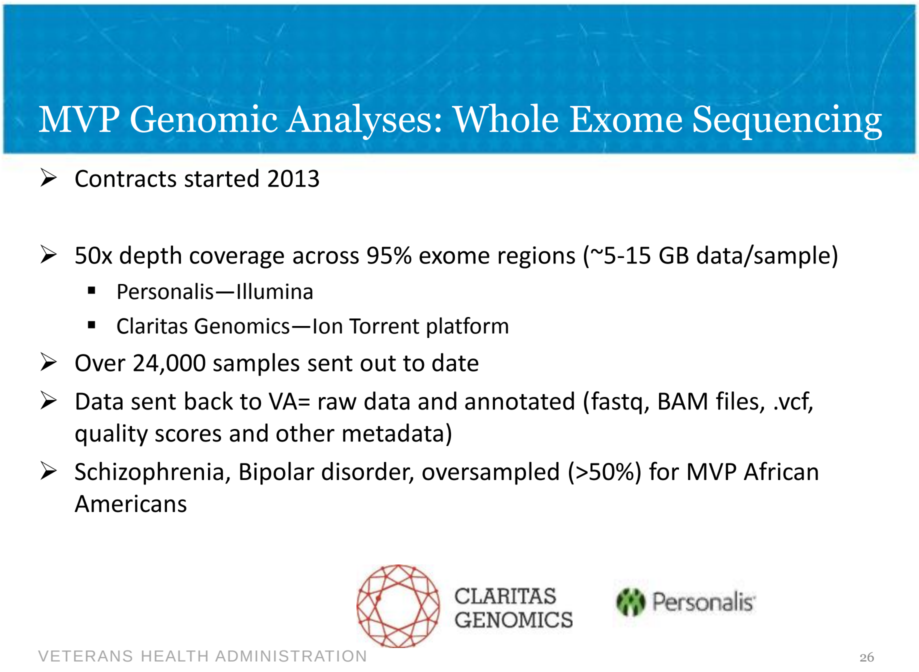


MVP Genomic Analyses: Whole Genome Sequencing

- Contracts started 2012
- 30x depth coverage over 90% genome (~300GB data/sample)
  - Personalis—Illumina platform
  - Claritas Genomics—Ion Torrent platform
- Almost 2,000 samples sent out to date
- Data sent back to VA as raw data and annotated (fastq, BAM files, .vcf, quality scores, some additional metadata)
- Samples = ALS, Schizophrenia, Bipolar disorder, and Exceptionally aged MVPers (95+)

CLARITAS GENOMICS

Personalis

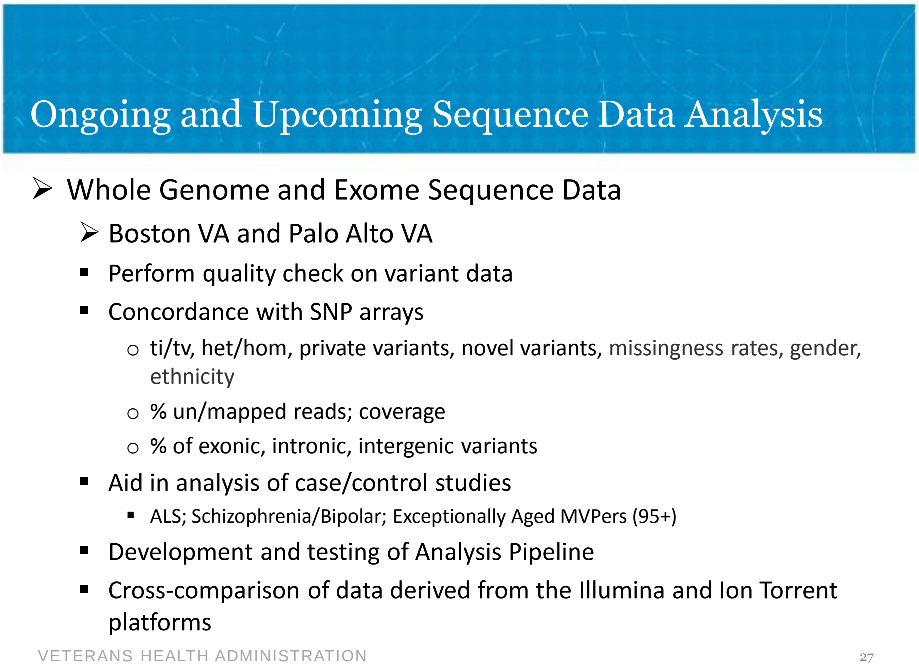


MVP Genomic Analyses: Whole Exome Sequencing

- Contracts started 2013
- 50x depth coverage across 95% exome regions (~5-15 GB data/sample)
  - Personalis—Illumina
  - Claritas Genomics—Ion Torrent platform
- Over 24,000 samples sent out to date
- Data sent back to VA= raw data and annotated (fastq, BAM files, .vcf, quality scores and other metadata)
- Schizophrenia, Bipolar disorder, oversampled (>50%) for MVP African Americans

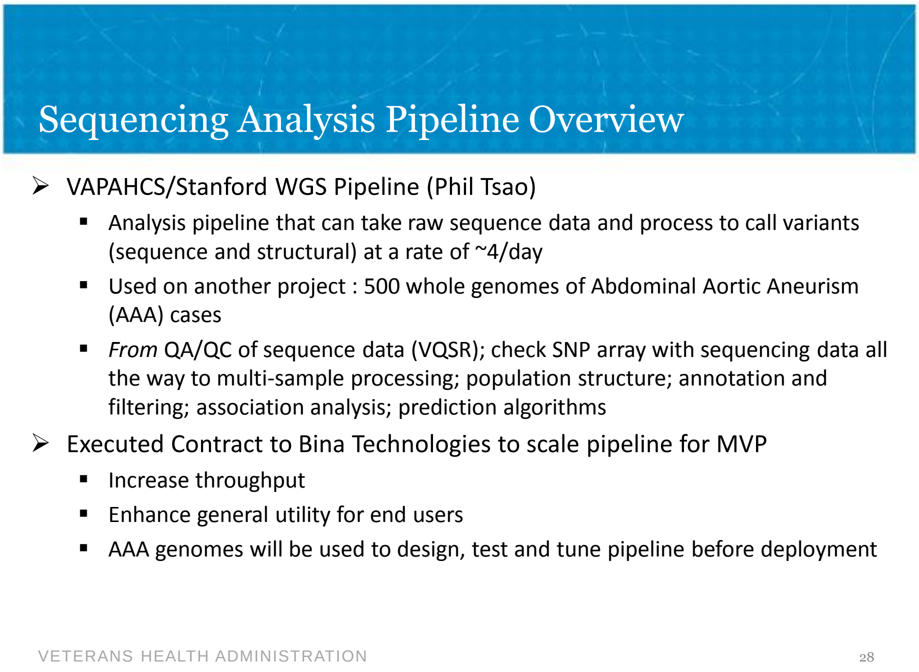
CLARITAS GENOMICS

Personalis



Ongoing and Upcoming Sequence Data Analysis

- Whole Genome and Exome Sequence Data
  - Boston VA and Palo Alto VA
  - Perform quality check on variant data
  - Concordance with SNP arrays
    - ti/tv, het/hom, private variants, novel variants, missingness rates, gender, ethnicity
    - % un/mapped reads; coverage
    - % of exonic, intronic, intergenic variants
  - Aid in analysis of case/control studies
    - ALS; Schizophrenia/Bipolar; Exceptionally Aged MVPers (95+)
  - Development and testing of Analysis Pipeline
  - Cross-comparison of data derived from the Illumina and Ion Torrent platforms



Sequencing Analysis Pipeline Overview

- VAPAHCS/Stanford WGS Pipeline (Phil Tsao)
  - Analysis pipeline that can take raw sequence data and process to call variants (sequence and structural) at a rate of ~4/day
  - Used on another project : 500 whole genomes of Abdominal Aortic Aneurism (AAA) cases
  - *From* QA/QC of sequence data (VQSR); check SNP array with sequencing data all the way to multi-sample processing; population structure; annotation and filtering; association analysis; prediction algorithms
- Executed Contract to Bina Technologies to scale pipeline for MVP
  - Increase throughput
  - Enhance general utility for end users
  - AAA genomes will be used to design, test and tune pipeline before deployment

MVP Genomic Analyses: SNP Genotyping

- Contracts started 2013
- Most "bang for the buck"
- A few MB data per sample
- Goal: genotype every MVP sample
  - *harness the power of large n*
- Which chip?

Used by other large biobanks (UK, Kaiser)

Customized MVP Affymetrix Axiom® Biobank Array

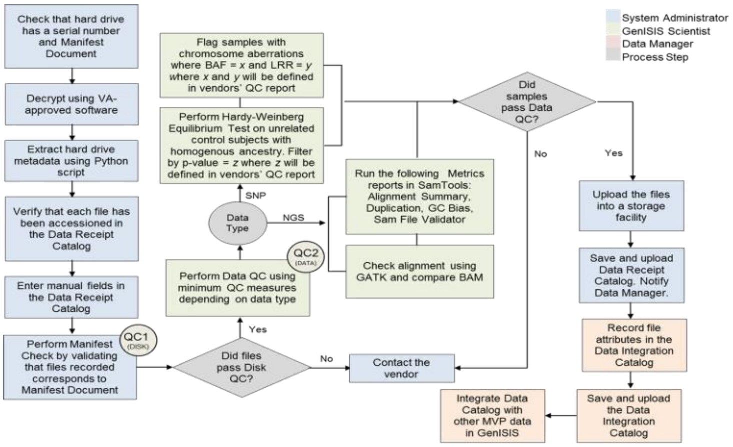
~723k features:

<u>BioBank Base Content - Revised</u>
Original Exome & Indel (238K)
Pharmacogenomic/ADME (2K)
eQTLs (23K)
New Exome & Indels (26K)
New LOF & Indels (70K)
GWAS – Published (246K)
GWAS – AA booster (50K)

<u>MVP Modules Added</u>
HLA/KIR (9k)
Psychiatric (26K)
Other disease/condition (42K) (Immuno, Cardio, Cancer, Blood types, Diabetic, ApoE, Addiction, Nephrology, Obesity, Stroke, Asthma, etc.)

Genomic Data Uptake Process

- All data shipped via hard drives (for now)



System Administrator
GenISIS Scientist
Data Manager
Process Step

Check that hard drive has a serial number and Manifest Document

Decrypt using VA-approved software

Extract hard drive metadata using Python script

Verify that each file has been accessioned in the Data Receipt Catalog

Enter manual fields in the Data Receipt Catalog

Perform Manifest Check by validating that files recorded corresponds to Manifest Document

QC1 (DISK)

Did files pass Disk QC?

No

Contact the vendor

Yes

Perform Data QC using minimum QC measures depending on data type

QC2 (DATA)

Data Type

SNP

Flag samples with chromosome aberrations where BAF = *x* and LRR = *y* where *x* and *y* will be defined in vendors' QC report

Perform Hardy-Weinberg Equilibrium Test on unrelated control subjects with homogenous ancestry. Filter by p-value = *z* where *z* will be defined in vendors' QC report

NGS

Run the following Metrics reports in SamTools: Alignment Summary, Duplication, GC Bias, Sam File Validator

Check alignment using GATK and compare BAM

Did samples pass Data QC?

No

Yes

Upload the files into a storage facility

Save and upload Data Receipt Catalog. Notify Data Manager.

Record file attributes in the Data Integration Catalog

Save and upload the Data Integration Catalog

Integrate Data Catalog with other MVP data in GenISIS

MVP SNP Genotype Data Efforts Underway

Genotyping Data

- Complete Data QC for all genotype data from FY14
- Upload all FY14 Genotype data to Pittsburgh (HPC cluster)
- Group effort for Imputation of SNP data against1000 genome
- data integration
  - analysis environment (HPC) in Pittsburgh
    - clinical data, survey data, genotype data
- Data and application support for RFA

The Numbers

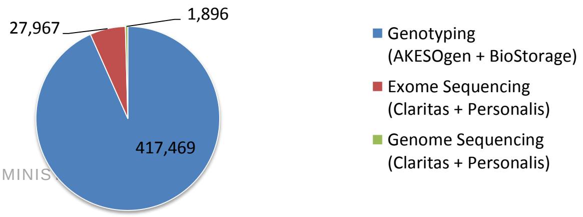
Sample Breakout by Analysis Type (FY12-14)

	SNP Genotype	Exome	Genome
MVP	408,113	18,611	612
CSP 572 (functional disability of SZ and BPI) cases	9,356	9,356	285
CSP 575* (PTSD)	*52,238*	*2,366*	N/A
ALS			999

*estimated number of PTSD cases from MVP cohort, based on self-reported survey data

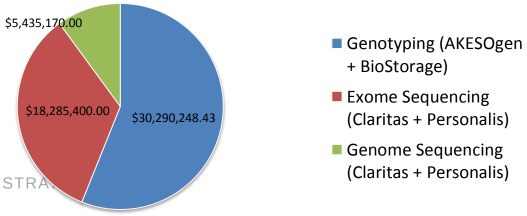
Analysis Types by Fiscal Year

	FY12	FY13	FY14	Totals*
SNP Genotyping (AKESOgen + BioStorage)		206,303	211,166	417,469
Exome Sequencing (Claritas + Personalis)		24,260	3,707	27,967
Genome Sequencing (Claritas + Personalis)	1,370	516	10	1,896
				*not all unique due to planned overlap



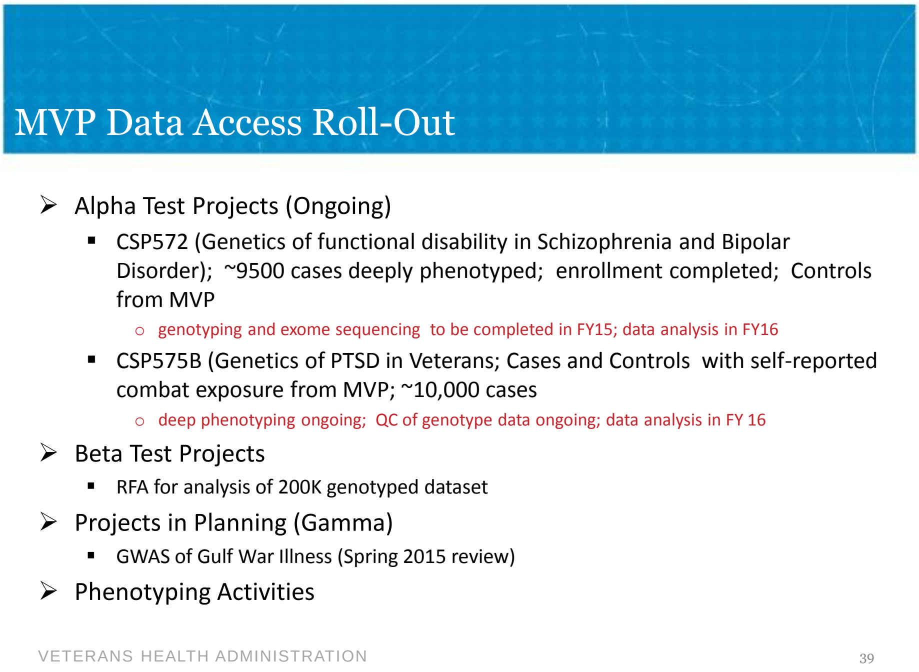
Contracting Budget by Fiscal Year

	FY12	FY13	FY14	Totals
Genotyping (AKESOgen + BioStorage)		$14,999,858.00	$15,290,390.43	$ 30,290,248.43
Exome Sequencing (Claritas + Personalis)		$16,798,200.00	$1,487,200.00	$ 18,285,400.00
Genome Sequencing (Claritas + Personalis)	$4,193,570.00	$1,200,780.00	$40,820.00	$ 5,435,170.00
TOTAL	$4,193,570.00	$ 32,998,838.00	$ 16,818,410.43	$ **54,010,818.43**



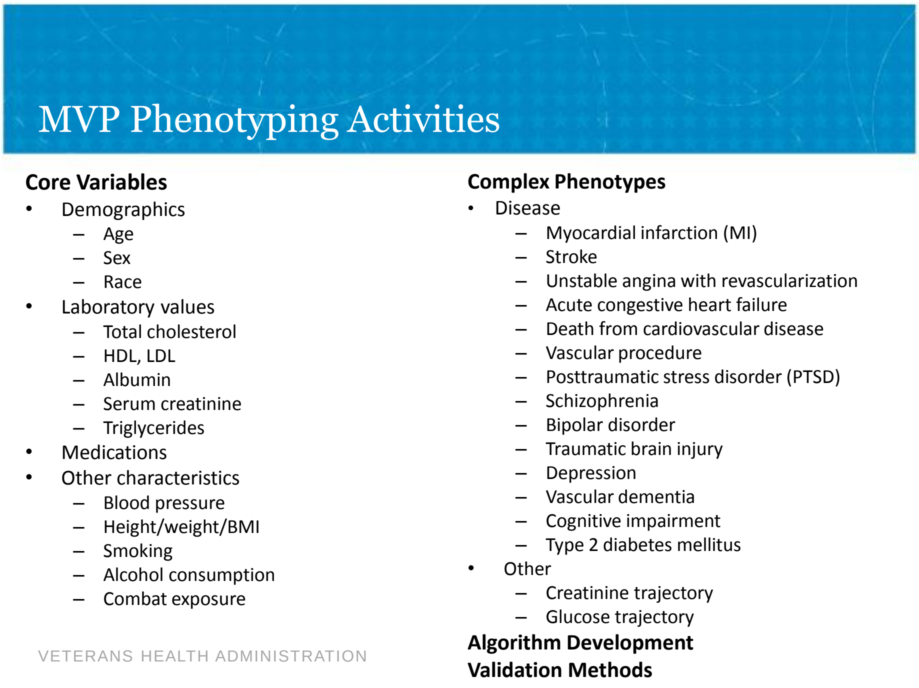
Data Access and Future Directions

- Framework to combine genetic data with phenotype data
- Access policy to allow approved researchers to utilize this combined data
- Future Directions



MVP Data Access Roll-Out

- Alpha Test Projects (Ongoing)
  - CSP572 (Genetics of functional disability in Schizophrenia and Bipolar Disorder); ~9500 cases deeply phenotyped; enrollment completed; Controls from MVP
    - genotyping and exome sequencing to be completed in FY15; data analysis in FY16
  - CSP575B (Genetics of PTSD in Veterans; Cases and Controls with self-reported combat exposure from MVP; ~10,000 cases
    - deep phenotyping ongoing; QC of genotype data ongoing; data analysis in FY 16
- Beta Test Projects
  - RFA for analysis of 200K genotyped dataset
- Projects in Planning (Gamma)
  - GWAS of Gulf War Illness (Spring 2015 review)
- Phenotyping Activities



MVP Phenotyping Activities

Core Variables

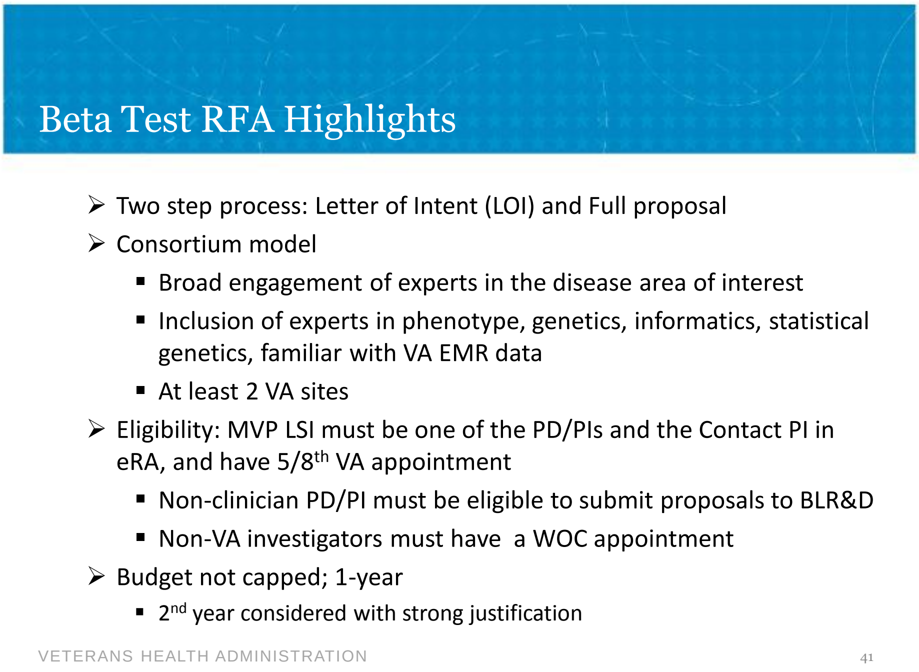
- Demographics
  - Age
  - Sex
  - Race
- Laboratory values
  - Total cholesterol
  - HDL, LDL
  - Albumin
  - Serum creatinine
  - Triglycerides
- Medications
- Other characteristics
  - Blood pressure
  - Height/weight/BMI
  - Smoking
  - Alcohol consumption
  - Combat exposure

Complex Phenotypes

- Disease
  - Myocardial infarction (MI)
  - Stroke
  - Unstable angina with revascularization
  - Acute congestive heart failure
  - Death from cardiovascular disease
  - Vascular procedure
  - Posttraumatic stress disorder (PTSD)
  - Schizophrenia
  - Bipolar disorder
  - Traumatic brain injury
  - Depression
  - Vascular dementia
  - Cognitive impairment
  - Type 2 diabetes mellitus
- Other
  - Creatinine trajectory
  - Glucose trajectory

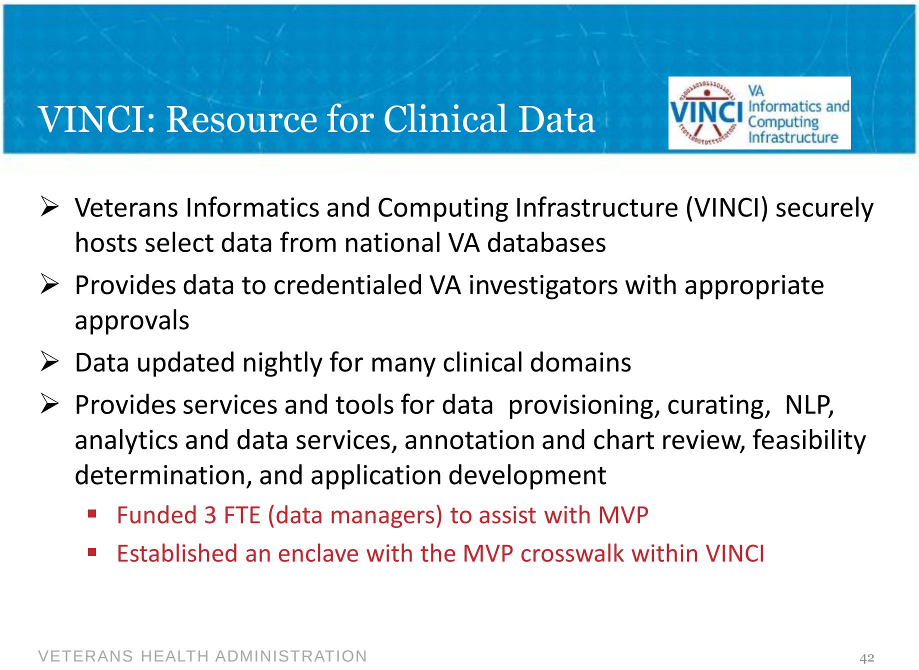
Algorithm Development

Validation Methods



Beta Test RFA Highlights

- Two step process: Letter of Intent (LOI) and Full proposal
- Consortium model
  - Broad engagement of experts in the disease area of interest
  - Inclusion of experts in phenotype, genetics, informatics, statistical genetics, familiar with VA EMR data
  - At least 2 VA sites
- Eligibility: MVP LSI must be one of the PD/PIs and the Contact PI in eRA, and have 5/8th VA appointment
  - Non-clinician PD/PI must be eligible to submit proposals to BLR&D
  - Non-VA investigators must have a WOC appointment
- Budget not capped; 1-year
  - 2nd year considered with strong justification



VINCI: Resource for Clinical Data

VINCI
VA Informatics and Computing Infrastructure

- Veterans Informatics and Computing Infrastructure (VINCI) securely hosts select data from national VA databases
- Provides data to credentialed VA investigators with appropriate approvals
- Data updated nightly for many clinical domains
- Provides services and tools for data provisioning, curating, NLP, analytics and data services, annotation and chart review, feasibility determination, and application development
  - Funded 3 FTE (data managers) to assist with MVP
  - Established an enclave with the MVP crosswalk within VINCI

GenISIS Secure Computing Environment

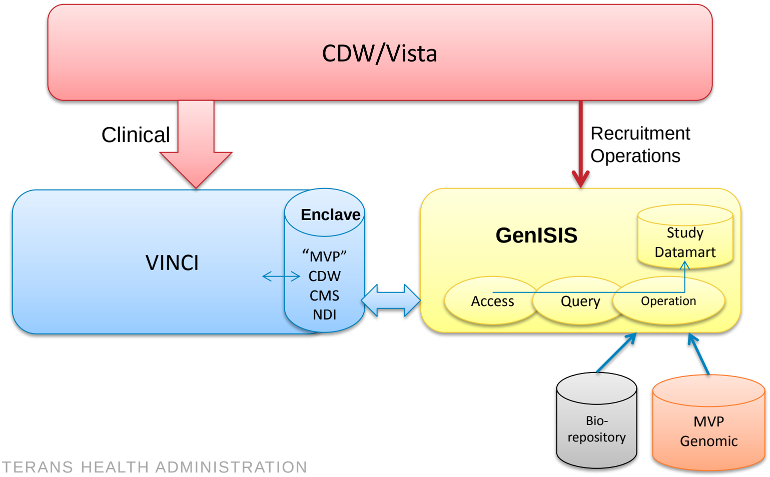
- Provides a secure computing environment with analytical tools
- Allows users to submit and manage their HPC jobs from a web dashboard
- Users authenticated by their VA network accounts
- Default 5 TB space allocated to each project, plus additional scratch space
- Analytic software available (PLINK, R/Bioconductor, SAS, Matlab, Perl/Bioperl, JMP Genomics and others) and will be updated based on user needs

Study Conduct

Following LOI approval and full proposal funding in Spring 2015:

- GenISIS Scientific Computing Environment (SCE)
  - Registration – valid VA NT and email accounts
  - Meet requirements of funding approval, completion of JIT requirements
- Genotypic data on ~200K samples
- Study-specific clinical data set imported from VINCI following DART approval
- MVP baseline and Lifestyle Survey data (if applicable)
- Analytic tools for genotype-phenotype association
  - Can import custom tools into the SCE

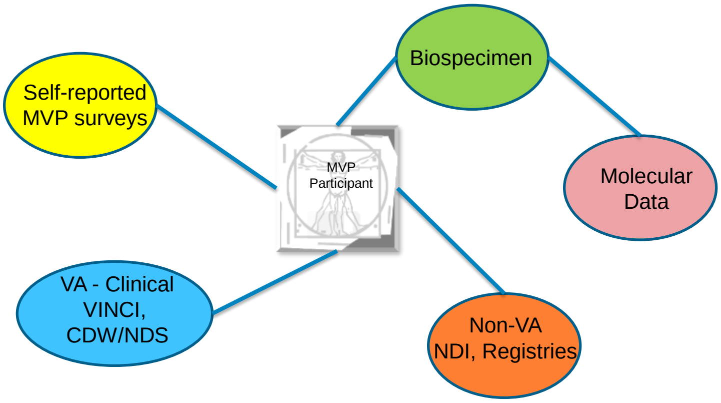
VINCI-GenISIS Convergence



Timeline for Beta Test Projects

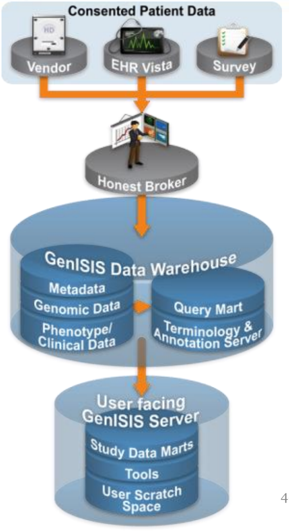
- RFA and Guidance Release: September 15, 2014
- Deadline for LOI Submission: November 15, 2014
  - 30 LOIs received
- Announcement of LOI Approval: December 15, 2014
  - 20 LOIs approved
- Proposal Submission Deadline in Grants.gov: March 10, 2015
- Scientific Review: June 10, 2015
- Funding Announcement: July 2015

MVP Data Universe

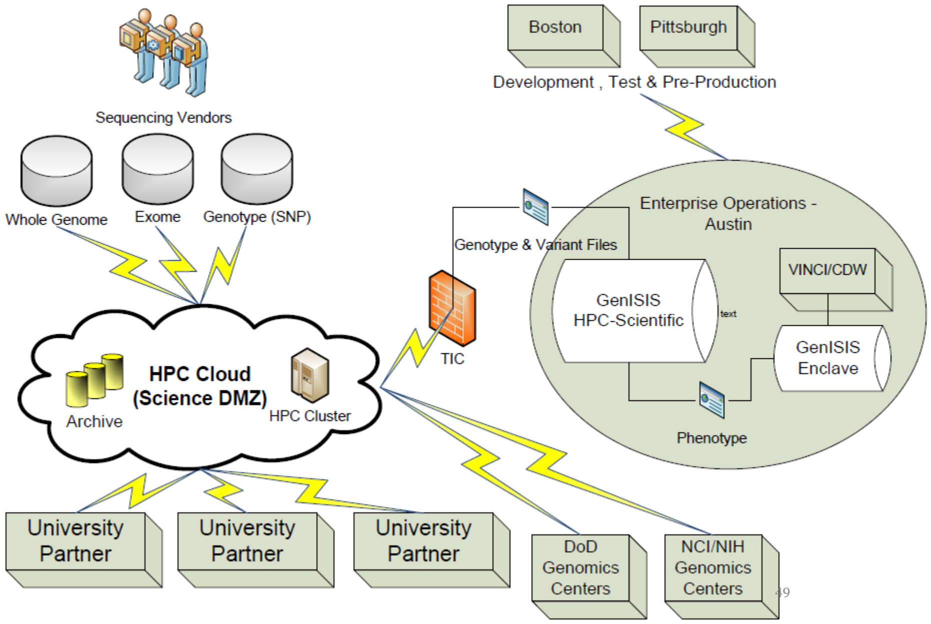


MVP Data Storage, Annotation and Integration

- Metadata extracted from clinical, survey and genomic data and their QA/QC metrics will be cataloged in a ***Metadata Database***
- Genomic data will be linked with corresponding clinical and survey data by an ***Honest Broker*** system
- ***Terminology and Annotation Server*** will allow researchers to incorporate a wide array of genomic and clinical annotations to integrated genomic, survey, and clinical data
- ***Query Mart*** will enable researchers to build cohorts and subset data using clinical and genomic information before exporting to the ***Study Data Mart*** for analysis



MVP “To-Be” Architecture



Future Plans for Data Access Roll -Out

- FY16 - RFA for expanded access to all VA medical centers (August 2015)
  - Limit of one LOI per site
  - Continue consortium model
  - Full proposal review in Spring 2016
  - Genotype data on 400K samples (underway)

- FY17 – RFA for expanded access to all ORD services
  - Epidemiological and GWAS
  - Pilot access to exome and genome sequence data
  - Pilot access to non-VA researchers (analysis within the GenISIS SCE – could be accelerated into FY 2016)

MVP Future Directions: Alternative Enrollment Strategies

- Develop user-platform/device-independent, web-based framework for survey completion, enrollment (Informed Consent + HIPAA Authorization), updating of information, and self-scheduling of study visits
  - Integration with other VA systems (ex. Identity and Access Management for identity authentication; MyHealtheVet)
- Explore alternate sample/collection strategies to with online enrollment
  - Pilot Saliva or buccal swab
  - Pilot mail consent
- Pilot enrollment at offsite, non-VA events
  - Mobile Vet Center bus

MVP Future Directions (Continued)

- Expansion of IT/Informatics infrastructure for storage and computation of large amounts of genomic data (FY16-17)
  - Austin data center
  - Cloud storage/computing
- Pilot MVP 2.0 - transitioning into clinical applications (FY 16-17)
  - CLIA sample collection, processing and analysis
  - Return of genetic test results (per ACMG guidelines)
  - Clinical decision support
  - Delivering personalized medicine to Veterans

"Our statistician will drop in and explain why you have nothing to worry about."