Common Controls

Steve Buyske
Rutgers University
GSPCC

Genome Sequencing Program Steering Committee Teleconference
November 17, 2107

Current
Case/Control Numbers

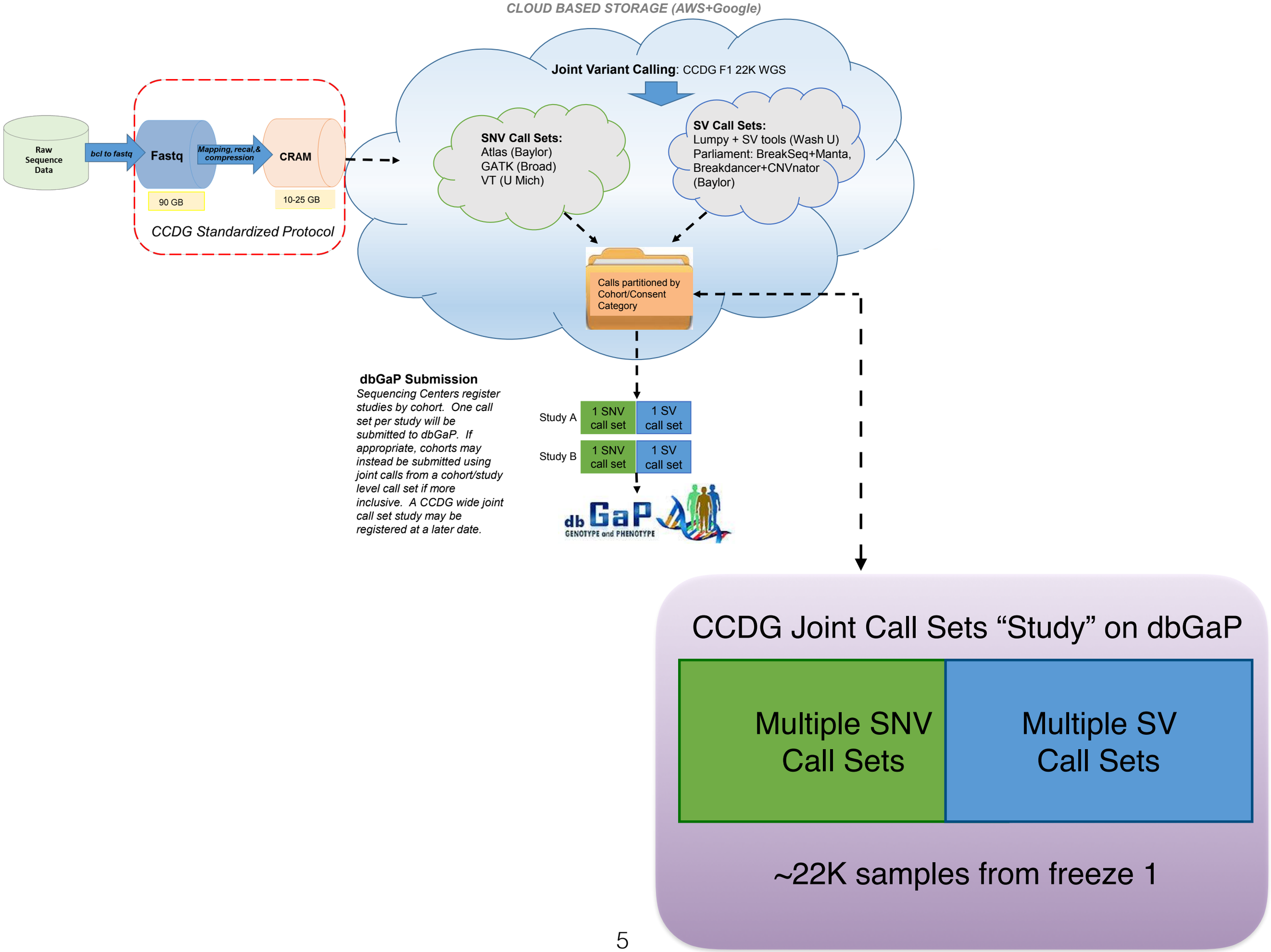
Disease	Cases			Controls		
	WES/WGG	WGS	Total	WES/WGG	WGS	Total
Autoimmune						
Asthma		650	650		521	521
IBD	848	3,320	4,168	600	332	932
T1D		1,397	1,397		1,507	1,507
Cardiovascular Disease						
Atrial Fibrillation		3,218	3,218		0	0
Coronary Artery Disease	8,429	8,217	16,646	6,383	6,757	13,140
Stroke	750	472	1,222	750	250	1,000
Multiple CVD study controls		0	0		7,797	7,797
Neuropsychiatric						
Alzheimer's		418	418		922	922
Autism	2,357	2,512	4,869	4,712	6,197	10,909
Epilepsy	13,501	0	13,501	412	0	412
Total	**25,885**	**20,204**	**46,089**	**12,857**	**24,283**	**37,140**

Target (Y3 Cumulative) Case/Control Numbers

Disease	Cases			Controls		
	WES/WGG	WGS	Total	WES/WGG	WGS	Total
Autoimmune						
Asthma		660	660		526	526
IBD	6,681	4,153	10,834	3,517	749	4,266
T1D		2,830	2,830		2,175	2,175
Cardiovascular Disease						
Atrial Fibrillation		5,450	5,450		800	800
Coronary Artery Disease	15,929	13,404	29,333	13,883	9,944	23,827
Stroke	750	3,478	4,228	750	250	1,000
Multiple CVD study controls		0	0		10,586	10,586
Neuropsychiatric						
Alzheimer's		1,100	1,100		1,100	1,100
Autism	3,157	5,680	8,837	6,312	7,960	14,272
Epilepsy	18,501	0	18,501	412	0	412
Total	**45,018**	**36,755**	**81,773**	**24,874**	**34,090**	**58,964**

Community Access

- Plans moving along for a dbGaP “Study” of the CCDG joint call sets
- Initially about 200 TB of VCFs for the ~22,000 samples in the Freeze 1 call sets
- Useful for
  - common controls repository
  - population genetics work
  - methods comparison
- Not something that could be recreated from individual studies’ dbGaP submissions


CLOUD BASED STORAGE (AWS+Google)
Raw Sequence Data
bcl to fastq
Fastq
90 GB
Mapping, recal,& compression
CRAM
10-25 GB
CCDG Standardized Protocol
Joint Variant Calling: CCDG F1 22K WGS
SNV Call Sets:
Atlas (Baylor)
GATK (Broad)
VT (U Mich)
SV Call Sets:
Lumpy + SV tools (Wash U)
Parliament: BreakSeq+Manta, Breakdancer+CNVnator (Baylor)
Calls partitioned by Cohort/Consent Category
dbGaP Submission
Sequencing Centers register studies by cohort. One call set per study will be submitted to dbGaP. If appropriate, cohorts may instead be submitted using joint calls from a cohort/study level call set if more inclusive. A CCDG wide joint call set study may be registered at a later date.
Study A
1 SNV call set
1 SV call set
Study B
1 SNV call set
1 SV call set
db GaP
GENOTYPE and PHENOTYPE
CCDG Joint Call Sets “Study” on dbGaP
Multiple SNV Call Sets
Multiple SV Call Sets
~22K samples from freeze 1

dbGaP CCDG Joint Call Sets Study

- Consents:
  - GRU: 4,695 HMB: 5,418 DS: 11,860
- Data:
  - Multiple SNV call sets
  - Multiple SV call sets
  - QC metrics
  - Possibly principal components of ancestry
  - Case/control status (and for which disorder)
  - Other phenotypes . . .

Wider Collection of Phenotypes in CCDG

- The trait working groups have case/control status and necessary covariates

- The originating studies have additional phenotypes, but may not have been expecting to share them with the CCDG (other than via mechanisms like dbGaP)

- Cohorts such as ARIC and SoL/HCHS already have phenotypes in dbGaP

- Pulling more phenotypes into CCDG would require additional effort

Some Analytic Issues

- Common controls might differ from study controls:
  - ancestry and other sources of subject heterogeneity
  - sequencing center
  - sequencing platform (standardized within GSP)
  - read depth
  - alignment pipeline (standardized within GSP)
  - variant calling algorithm

Some Analytic Issues

- All of the potential differences (on previous slide) are potential pitfalls in analysis

- The Common Controls WG is putting together a review of relevant analytic approaches that address these issues

First Paper in Preparation

- Led by Chris DeBoever (Stanford)
- Focus: rare variant associations, multi-ancestry, and empirical power estimates
- Risk/protective differences, effect of disease prevalence
- Power to detect LoF associations in different populations based on allele frequency estimates
- Multi-ancestry genome-wide association simulations (coalescent-based)
  - What is the probability of observing a rare variant association across populations?
  - Is local ancestry estimation useful for rare variant associations with admixed subjects?