CCDG Data Flow Working Group

Co-chairs

Ginger Metcalf (HGSC) Steve Buyske (GSPCC)

November 17, 2017



National Human Genome Research Institute

- Identify CCDG Deliverables and Recipients
- Determine Strategy for Delivering Data to CCDG WG, GSP Investigators and Public
- Work with other CCDG WG to determine data freeze schedule
- Collect and Organize consent information, sample metadata, and QC Metrics
- Identify Long Term Storage Organization (dbGaP)

CLOUD BASED STORAGE (AWS+Google) Joint Variant Calling: CCDG F1 22K WGS SV Call Sets: SNV Call Sets: Lumpy + SV tools (Wash U) Mapping, recal,& compression Raw Atlas (Baylor) Parliament: BreakSeq+Manta, bcl to fastq Fastq CRAM Sequence GATK (Broad) Breakdancer+CNVnator Data VT (U Mich) (Baylor) 10-25 GB 90 GB CCDG Standardized Protocol Calls partitioned by Cohort/Consent Category **GSP** Dissemination Callsets accessed directly from the cloud, at the cohort level based on data use dbGaP Submission limitations. GSP CC Sequencing Centers register manages access to cloud studies by cohort. One call bucket, and tracks data use 1 SV 1 SNV Study A set per study will be parameters by cohort call set call set submitted to dbGaP. If appropriate, cohorts may 1 SV 1 SNV Study B instead be submitted using call set call set 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. **GENOTYPE and PHENOT**