Data Working Group Progress Report

NHGRI Genome Sequencing Program Meeting

April 12, 2016

Data Working Group Goals

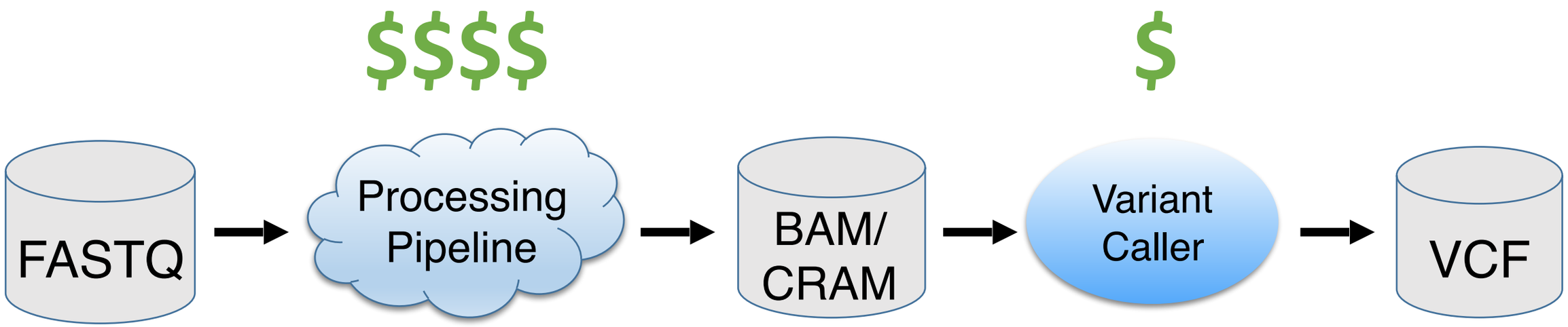
(1) Develop an efficient strategy for data sharing among consortium members

(2) Lead / participate in efforts to generate cross-center and cross-project genome variation call sets

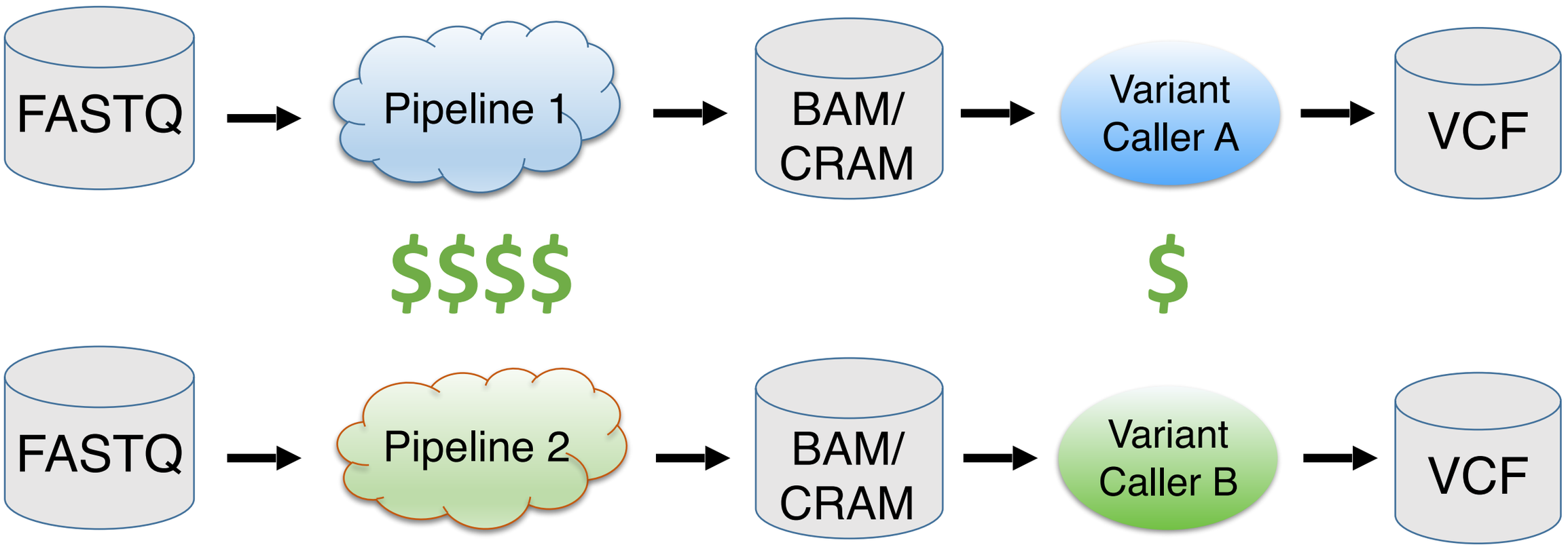
Efficient data sharing is crucial

- **CCDG will generate a large amount of data**
  - e.g., 100,000 genomes * 50 Gb/genome = 5 Petabytes
- **Collaborative projects will involve data produced at different centers**
- **Accurate variant calling requires *joint* analysis of raw (or nearly raw) read alignment data**
- **How do we put together variant call sets, quickly?**
  - How do we move data around?
  - How do we make data compatible to avoid reprocessing?

The typical data workflow



The problem

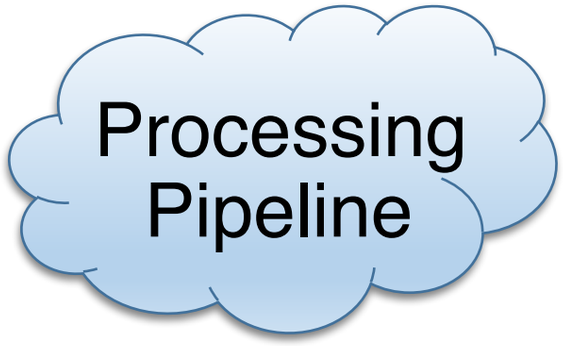


- **Comparing variant call sets generated by different groups is tricky**
  - Data processed at different sites is not generally compatible
  - Choices in reference genome, aligner, or data processing steps lead to different variant sites and genotypes
  - These data processing “batch effects” encumber analyses that seek to combine data from different projects or centers

Proposal: pipeline standardization

Guiding principles

- Make it possible to combine data from different centers
- Avoid need for expensive low-level reprocessing
- Retain some flexibility: allow pipeline efficiency improvements and variant caller innovation



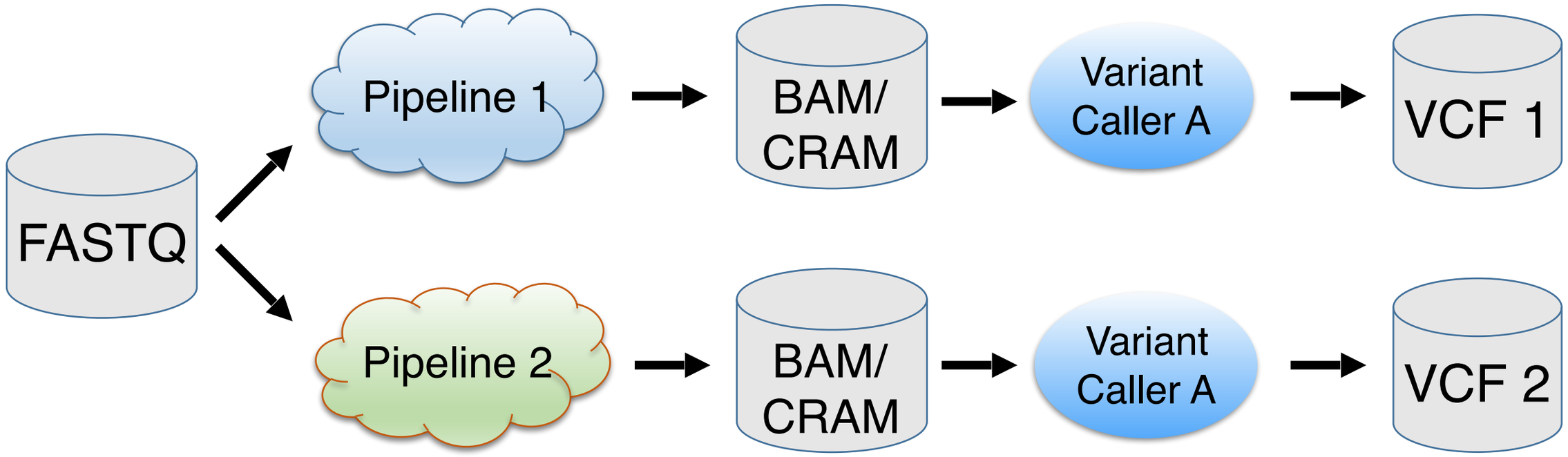
Our proposal aims to make alignment and data processing compatible

Our proposal assumes many variant callers can still be used with a set of alignments

Scope: Our current proposal spans CCDG & TOPMed. We recognize the need to engage with others.

When are two pipelines functionally equivalent?

- They can receive the same reads as input
- They can produce a BAM/CRAM file as output
- Running a variant caller on the two outputs produces nearly identical variants and genotypes



Processes satisfy functional equivalence if VCFs are “nearly identical”

What is needed to achieve pipeline standardization across centers?

(1) Define a standard

(2) Define datasets and metrics for testing

(3) Test and modify pipelines at each center until functional equivalence metrics are met

(4) Agree on process and timeline for future updates

Pipeline standardization effort: decision points and progress to date

- **Reference genome: GRCh38DH, 1KGP version**
- **Alignment software and parameters: nearly done**
  - BWA-MEM; agreement on parameters; seeking minor mods.
- **Duplicate marking: nearly done**
  - 3 tools; tentative agreement on standard; currently testing
- **Base quality score recalibration: nearly done**
  - 2 tools; tentative agreement on standard; currently testing
- **Base quality score compression: nearly done**
  - 3-bin (2,10,30) or 4-bin (2,10,20,30); currently testing
- **Indel realignment: remove**
- **Alignment file format: lossless CRAM**

*Decisions reached by consensus. We aimed to minimize file size & compute cost, adhere to SAM spec. & current best practices. Evidence-based conflict resolution.

Data resources to generate and share

- **Tier1: testing for initial pipeline standardization effort**
  - 5 x 1000 Genomes samples (including NA12878)
  - HiSeqX, diverse data quality
- **Tier2: long term effort to implement improved data processing and variant calling methods**
  - Diverse set of trios; maximize overlap with 1000 Genomes, Reference Genome Projects & Genome in a Bottle
  - Haploid hydatidiform mole genomes (CHM1 & CHM13)
  - Initial proposal: 32 genomes: 10 trios, 2 moles.

Pathway for updates to this data standard

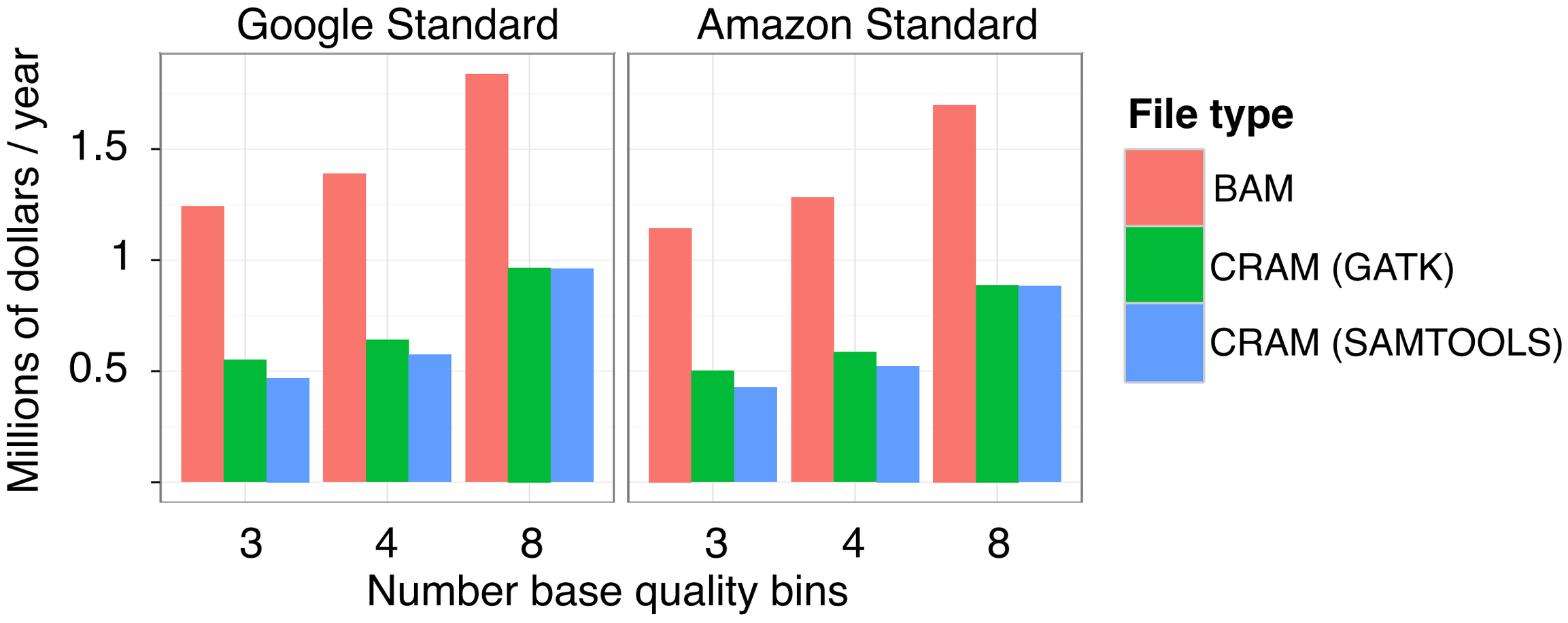
- Efficiency updates passing functional equivalence tests are always allowed
- In the future, we expect there will be better aligners, reference genomes, and data processing steps
- Propose to start a review process in late 2017: invite proposals for updates to improve variant calling

Ancillary benefits of pipeline standardization

- Any group can run a validated pipeline and accurately compare results with variant databases from CCDG or TOPMed
- Beyond alignment pipelines, these tests can also be applied to any long term data repository
- A long term repository must be able to receive an alignment and return a functionally equivalent alignment at a future time point

File size & data storage cost reductions

- Using BAM and 8-bin base quality scores, a typical 30X WGS dataset is 54 Gb = 5.4 Pb for 100K genomes
- A CRAM file using our proposed pipeline is 14-19 Gb = 1.4-1.9 Pb for 100K genomes (26-35%)



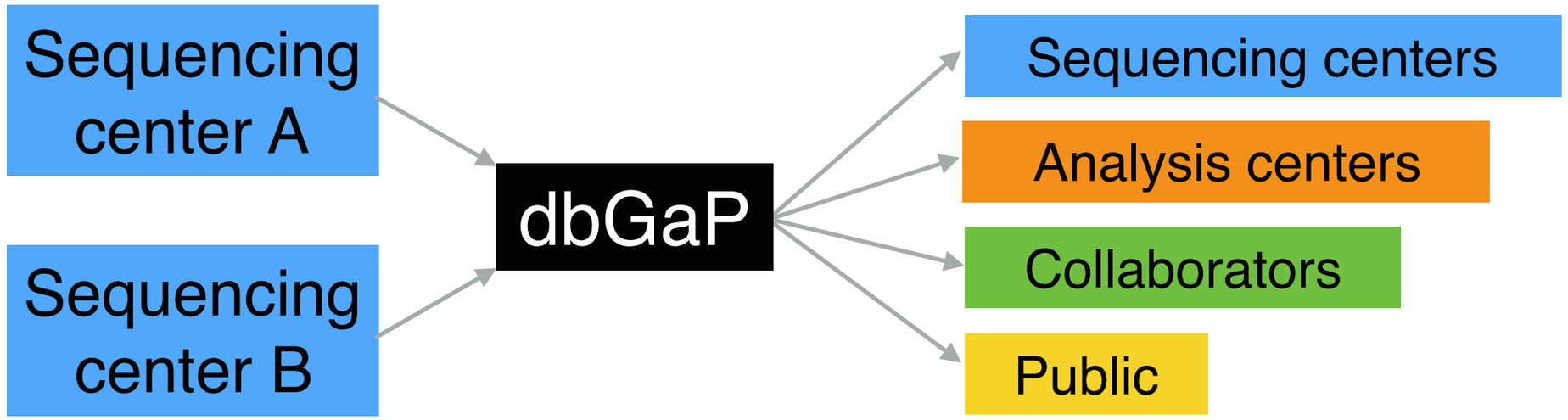
Two negative consequences of aggressive file size reduction

- **Loss of original base quality scores**
  - Do we care?
  - Should we store these elsewhere?
    - If so, where? SRA? Sequencing centers?
- **Labor; many minor software & pipeline updates required to work with CRAM**

<u>Question</u>: for any given project, how do we aggregate data at a single site for variant calling?

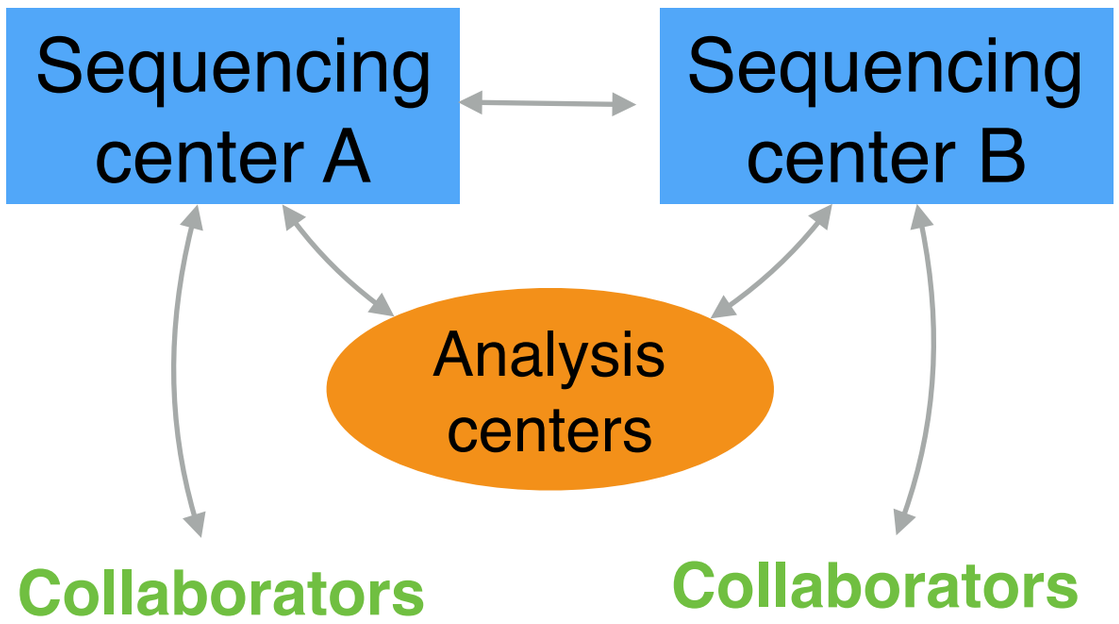
Data sharing mechanisms to consider

The current model:

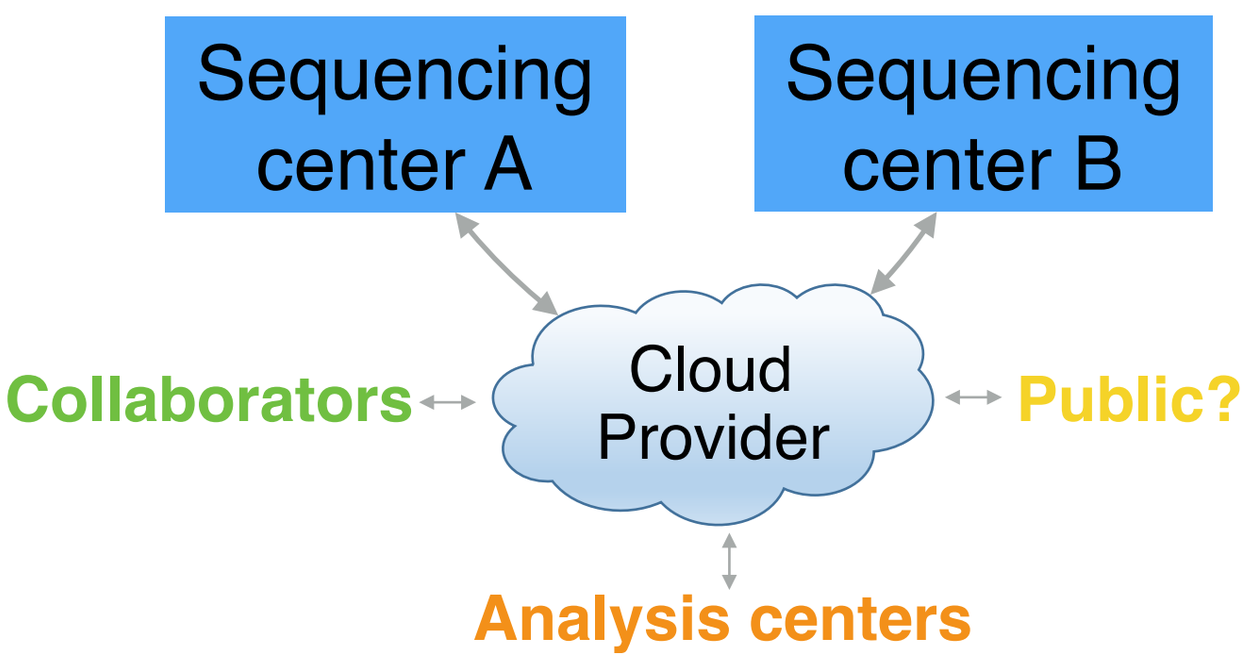


Two other potential models:

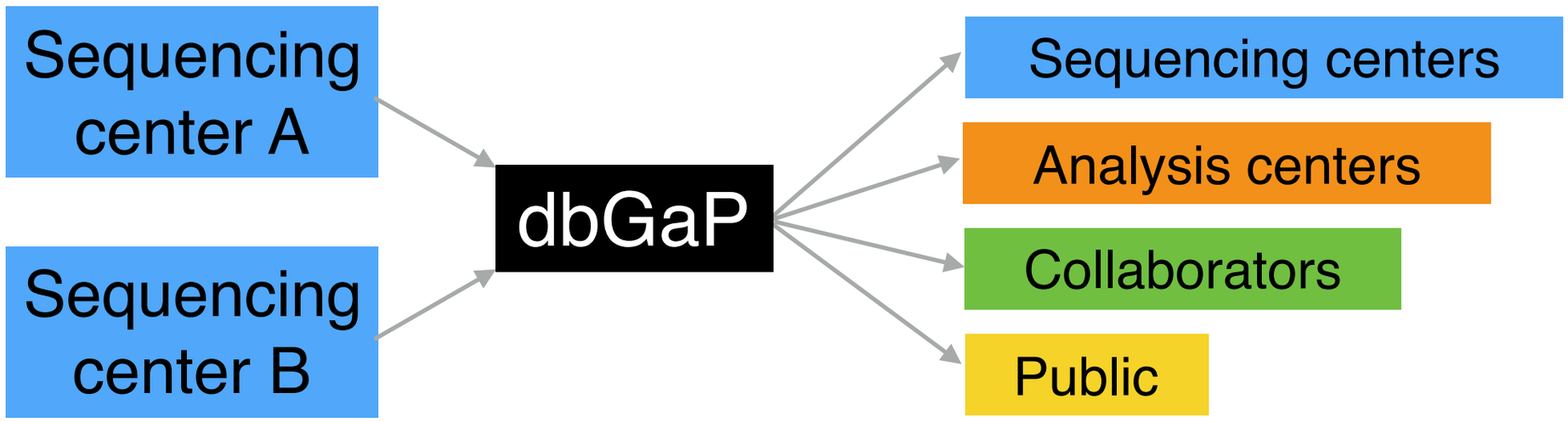
direct file transfer



cloud “sandbox”



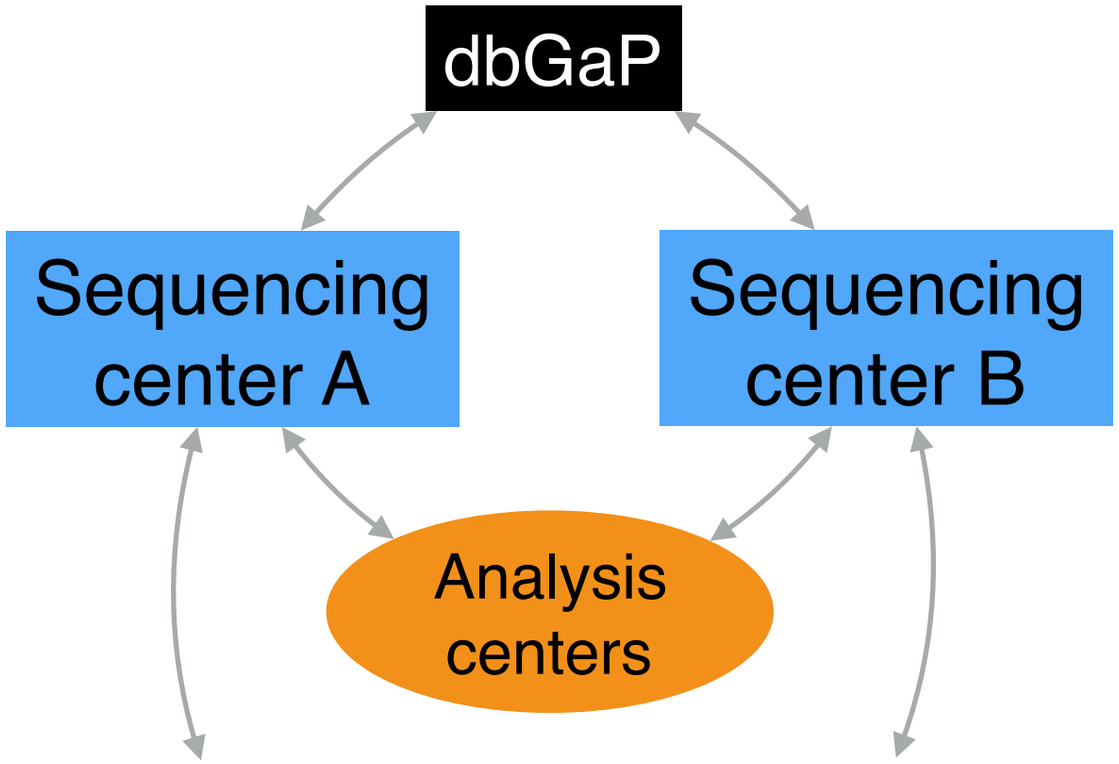
Data Sharing Recommendations



- We expect dbGaP to be the primary mechanism for long-term data storage and dissemination.
- However, we are concerned about using dbGaP as the sole means for data sharing within the consortium.

Data Sharing Recommendations

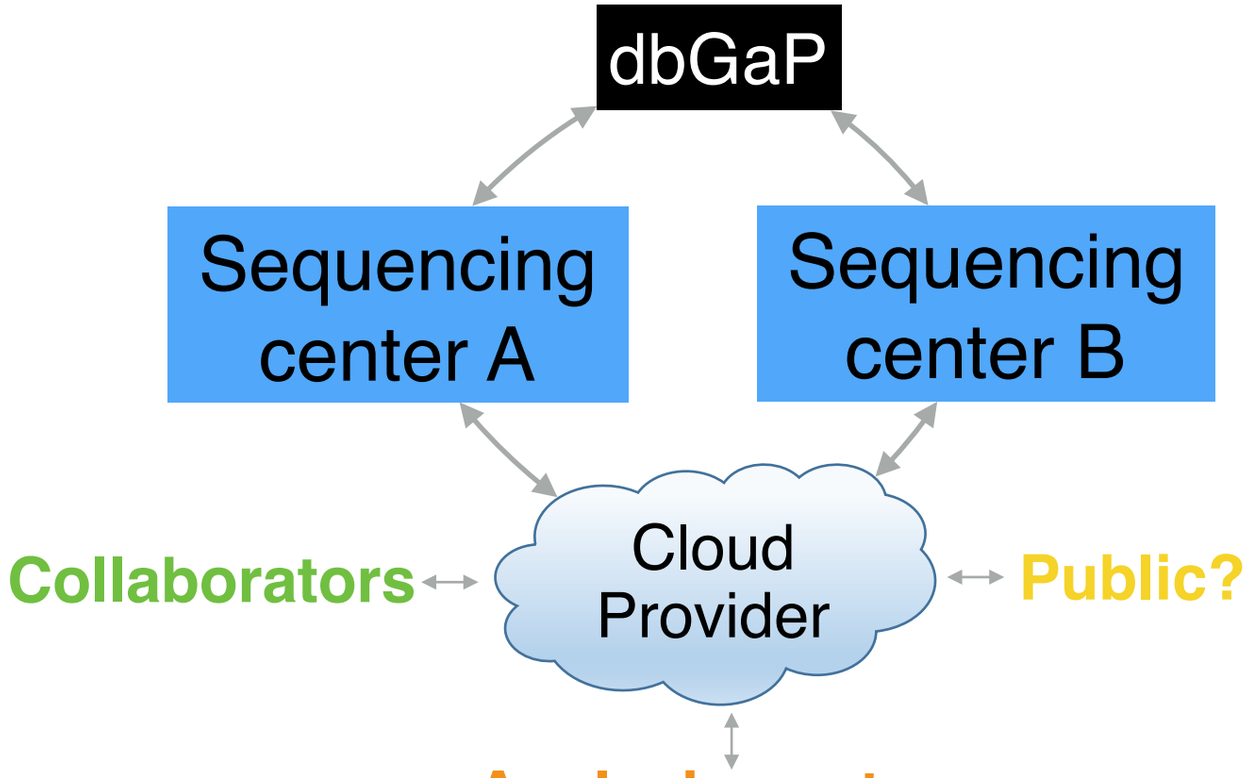
Direct file transfer will be the most efficient short-term model for sharing large datasets among consortium members



- Advantages:
  - Simple & fast: easy set-up; >1,000 genomes/day transfer
- Challenges:
  - Regulation: access control? data tracking? transparency?
  - Scope: will be difficult to serve many people; limit to centers?
  - Sustainability: this should be a temporary solution

Data Sharing Recommendations

Ideally, all CCDG data would be accessible on a cloud-based analysis "sandbox"



- Advantages:
  - Access: data readily accessible to all parties
  - Collective cost: minimizes file storage redundancy across sites
- Challenges:
  - Short term uncertainty: Which provider? access control?
  - Cost: more expensive for groups with ample local compute; data egress charges (~$150K for 100K genomes using Google).
- Our working group has not spent much time on this issue yet; more research and discussion is needed.

Data sharing caveats

- NIH policy
- Human subject consents
- Pre-existing data sharing agreements
- Institutional review boards

The Data WG did not tackle these issues.
We need help from the policy experts.

Future (near-term) plans

- Complete pipeline standardization effort – the clock is ticking
- Finalize and implement data sharing plan (in conjunction with NIH staff and policy experts)
- Move on to the fun stuff!

Data Working Group Members & Contributors

Core Members

Ira Hall (WashU)
Benjamin Neale (Broad)
Michael Zody (NYGC)
William Salerno (Baylor)
Steve Buyske (Rutgers)

TOPMed Representative

Goncalo Abecasis (U. Mich.)

Pipeline standardization

Hyun Min Kang (U. Mich.)
Dave Larson (WashU)
Allison Regier (WashU)
Eric Banks (Broad)
Yossi Farjoun (Broad)
Kathleen Tibbets (Broad)

Additional participants

Adam Felsenfeld (NHGRI)
Carolyn Hutter (NHGRI)
Heidi Sofia (NHGRI)
Cashell Jaquish (NHLBI)
Jinchuan Xing (Rutgers)
Tara Matise (Rutgers)

Supplemental Slides

A downside to cloud sharing: data egress