Update from the Data Analysis WG

Ira Hall, Ben Neale, Mike Zody,
Will Salerno, Goncalo
Abecasis, *et al*.

GSP Teleconference
Nov 17, 2017

A reminder of our goals

(1) Generate high-quality genome variation maps from combined GSP data

(2) Improve variant calling and annotation

- Use best possible methods (that we can afford)
- Know how well we are doing

(3) Share variant calls (Data Flow WG)

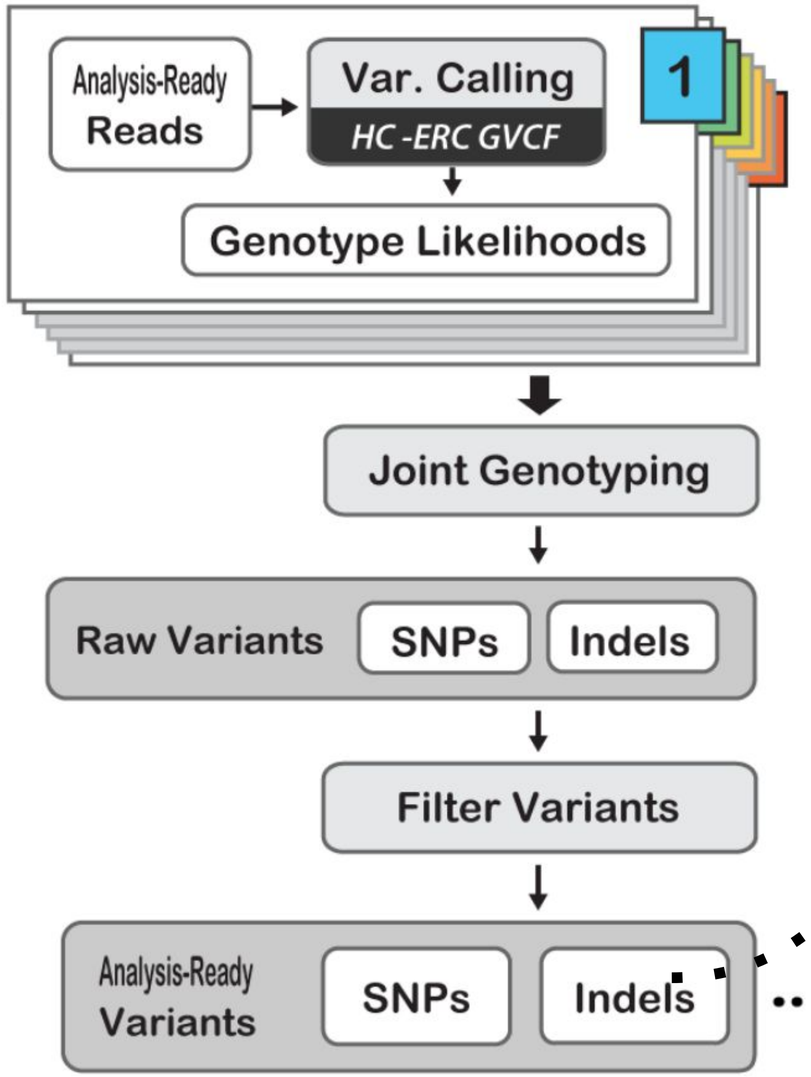
- Disease working groups
- Analysis Centers
- Research community

Activities and progress

- Pan-center pipeline harmonization and functional equivalence to enable data sharing
  - documentation published online: https://github.com/CCDG/Pipeline-Standardization/blob/master/PipelineStandard.md
  - manuscript in prep.
- WGS data aggregation on Google & Amazon
- Variant calling for CCDG Freeze 1
  - 22K genomes
  - Five callsets in various stages of completion
- Distribution scheme (Data Flow WG)
  - Ginger Metcalf, Tara Matise, etc.

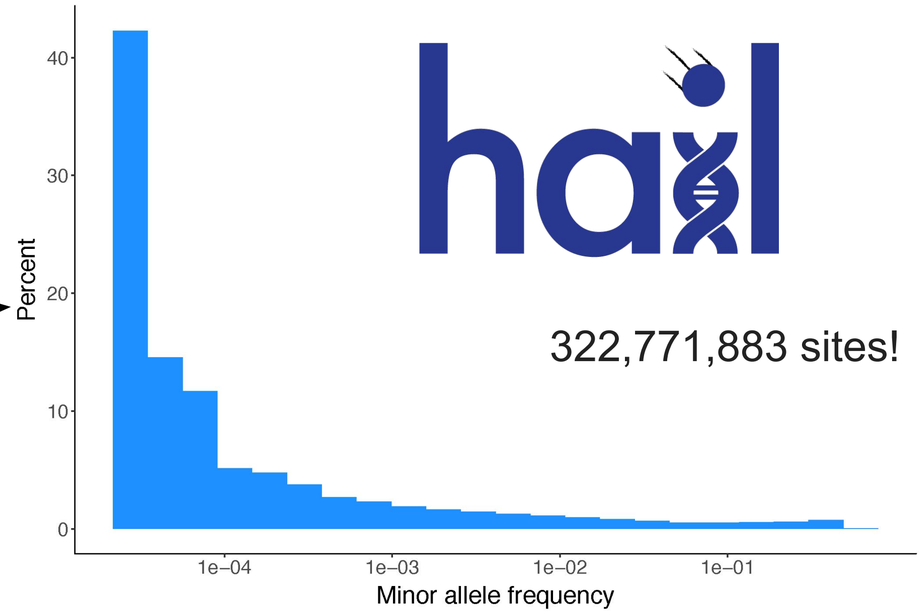
GATK Joint Calling + Hail QC

gatk VARIANT DISCOVERY



SNPs & Indels Jointly Called on 22K CCDG Samples

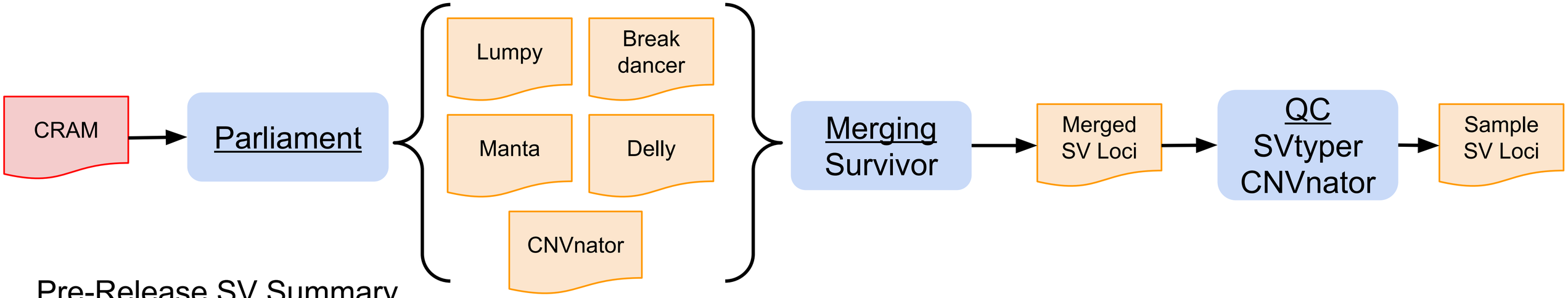
- Variant discovery and genotyping via GVCF pipeline in the GATK
- Standard GATK variant recalibration and site-level filtering
- Sample and genotype-level filtering and QC through Hail



hail

322,771,883 sites!

CCDG Freeze 1 Parliament Discovery

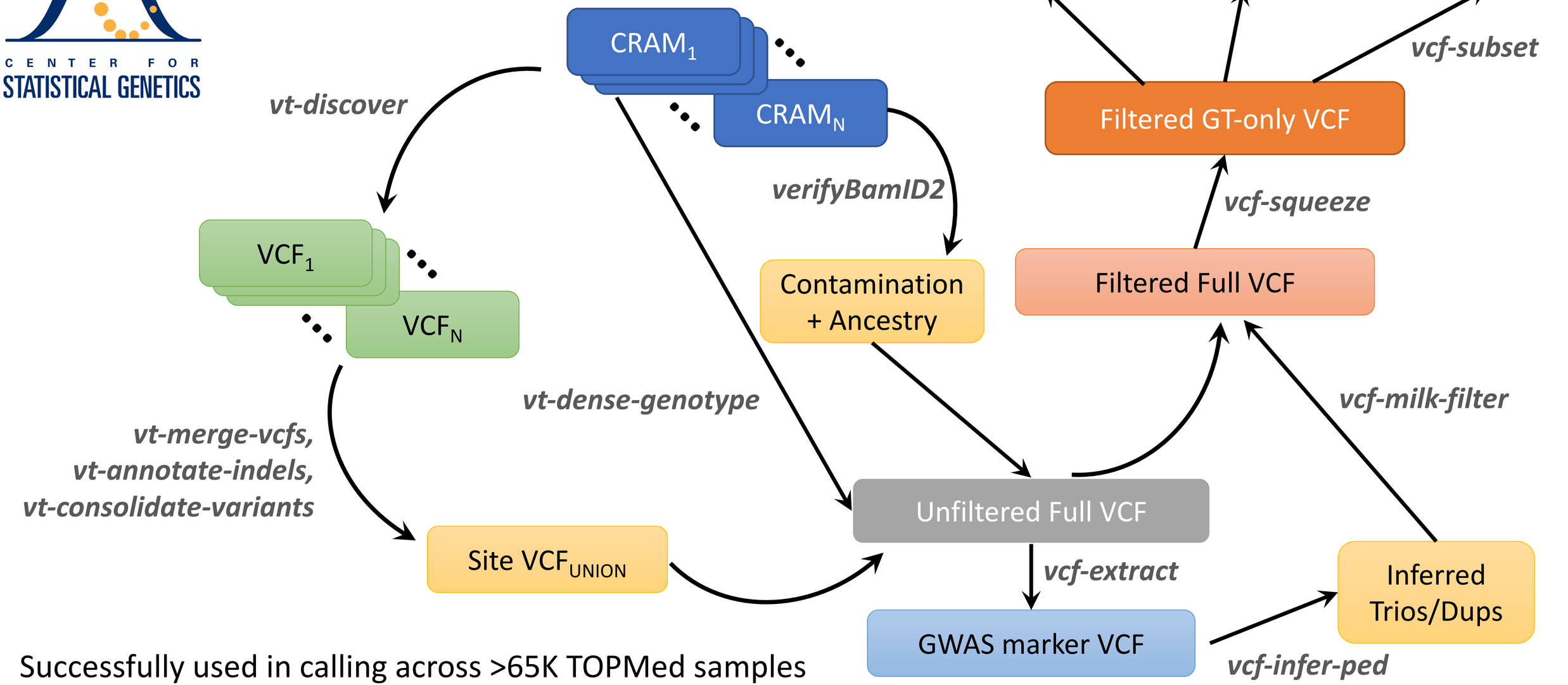


Pre-Release SV Summary

- Merge sample SV loci across Freeze 1
  - ~6.2 M Loci x 22,609 samples
- 11/17 Deliverables:
  - SV Loci list with Freeze 1 sample counts
  - Per-consent-group subsets
- Population genotyping to follow

#CHR	POS	ID	REF	ALT	INFO	FORMAT	Sample1	Sample2
1	1	DEL000SUR	N	<DEL>	C=332	GT:LN:DR:ST:TY:CO	.....:10000:0,0:+-:DEL:1_1-1_10000	.....:10000:0,0:+-:DEL:1_1-1_10000
1	92209	DEL00579SUR	N	<DEL>	C=2	GT:LN:DR:ST:TY:CO	.....:15:0,0:+-:DEL:1_92209-1_92224	./.:0:0,0:--:NaN:NaN
1	139701	DEL00792SUR	N	<DEL>	C=456	GT:LN:DR:ST:TY:CO	./.:0:0,0:--:NaN:NaN	./.:0:0,0:--:NaN:NaN
1	142101	DEL00836SUR	N	<DEL>	C=22001	GT:LN:DR:ST:TY:CO	./.:0:0,0:--:NaN:NaN	.....:13800:0,0:+-:DEL:1_141801-1_155600
1	206001	DEL001164SUR	N	<DEL>	C=1889	GT:LN:DR:ST:TY:CO	./.:0:0,0:--:NaN:NaN	.....:51600:0,0:+-:DEL:1_206201-1_257800
1	207301	DEL001167SUR	N	<DEL>	C=7764	GT:LN:DR:ST:TY:CO	.....:50400:0,0:+-:DEL:1_207301-1_257700	./.:0:0,0:--:NaN:NaN
1	297901	DEL001440SUR	N	<DEL>	C=78	GT:LN:DR:ST:TY:CO	.....:50600:0,0:+-:DEL:1_297801-1_348400	.....:50900:0,0:+-:DEL:1_297501-1_348400
1	385701	DEL001763SUR	N	<DEL>	C=675	GT:LN:DR:ST:TY:CO	./.:0:0,0:--:NaN:NaN	./.:0:0,0:--:NaN:NaN
1	388801	DEL001772SUR	N	<DEL>	C=1	GT:LN:DR:ST:TY:CO	.....:1200:0,0:+-:DEL:1_389101-1_390300	./.:0:0,0:--:NaN:NaN
1	393701	DEL001797SUR	N	<DEL>	C=103	GT:LN:DR:ST:TY:CO	.....:2500:0,0:+-:DEL:1_393601-1_396100	./.:0:0,0:--:NaN:NaN

GenomesOnTheCloud (GotCloud) Pipeline



Successfully used in calling across >65K TOPMed samples

WashU structural variation (SV) callset

MGI tools for population-scale SV mapping:

Layer et al., *Genome Biology* (2014)

Chiang et al., *Nature Methods* (2015)

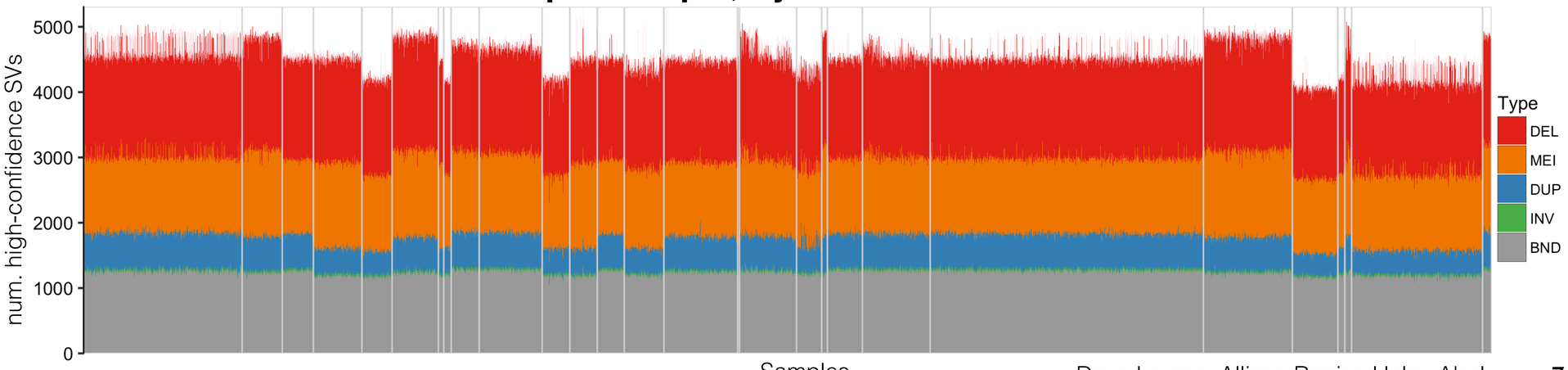
svtools: https://github.com/hall-lab/svtools

- Our methods:
  - SV discovery: LUMPY (Hall & Quinlan lab)
  - merging, classification, annotation, filtering: svtools (Hall lab)
  - breakpoint genotyping: svtools/svtyper (Hall lab)
  - copy number estimation: CNVnator (Abyzov & Gerstein labs)
  - efficient pipeline architecture from SpeedSeq (Hall lab)
- Scalability: ~23K deep WGS, Google Cloud, $0.65 / genome
- Tuned by mendelian error rate in family data
- Version 1 complete, ready to share (QC ongoing)

High confidence structural variants (n=356,948):

Type	Common	Low Freq.	Rare
DEL	4020	2389	148621
DUP	1089	437	38461
MEI	1969	263	3289
INV	52	31	1297
BND	1730	753	46056

SV counts per sample, by cohort & variant class:



Looking towards the future

- Characterize variant maps from each group
  - benchmarking: sensitivity, accuracy, efficiency, cost
  - genome biology; functional annotation
- New and improved variant calling methods
  - assembly to identify novel insertions (Zody et al., NYGC)
  - specialized genotyping for difficult variant classes
  - various other approaches
- Create community resources to aid gene mapping
  - variant servers; imputation panels; common controls
- Bigger, more informative datasets!
  - joint calling on future CCDG freezes
  - collaboration with other programs (TOPMed, WGSPD, etc.)

Key Contributors

CCDGs

Ira Hall (MGI)
Benjamin Neale (Broad)
Michael Zody (NYGC)
William Salerno (HGSC)
Allison Regier (MGI)
Yossi Farjoun (Broad)
Dave Larson (MGI)
Olga Krasheninina (HGSC)
Daniel Howrigan (Broad)
Eric Banks (Broad)

TOPMed IRC

Goncalo Abecasis (U. Mich.)
Hyun Min Kang (U. Mich.)

GSPCC

Tara Matise (Rutgers)
Jinchuan Xing (Rutgers)
Yeting Zhang (Rutgers)

Additional participants

Adam Felsenfeld (NHGRI)
Carolyn Hutter (NHGRI)
Heidi Sofia (NHGRI)
Cashell Jaquish (NHLBI)
Anjene Addington (NIMH)
Shane McCarthy (Sanger)
Kurt Hetrick (JHU CMG)
Josh Smith (UW CMG)
James Knight (Yale CMG)
Daniel MacArthur (Broad CMG)
Stacey Gabriel (Broad)
Bingshan Li (Vanderbilt AC)
Eimear Kenney (Mt. Sinai AC)
Li San Wang (Penn., ADSP)

*note: many other members not listed