1000G Phase 1 indel calling discussion

Mark DePristo
Genome Sequencing and Analysis
Medical and Population Genetics
March 8, 2011

Discussion points

- What indel call sets are currently available?
- What do we know about their quality?
- Are these call sets sufficiently good to release as is? Or do we need to devote additional resources to improve the methods?
- Validation
  - Which indels should we validate?
  - What technology should we use?

Data and definitions

- Evaluation data sets:
  - EUR chr20 call sets from GATK, DINDEL, and samtools
  - Union of all three
  - Control: GATK SNP calls for EUR+ samples, Project-consensus VQSR High-sensitivity
- Comparison data sets:
  - Complete genomics indel calls for 38 hapmap individuals
  - Homozygous SNP and indel sites in NA12878
    - Very unlikely to be errors
    - Complete genomics
    - Illumina HiSeq at 64x, called with the GATK
  - Pilot 1 SNP and indel validation sites from 1000G
- For technical reasons, I consider any call at the same left-aligned site in two data sets as the same

The indel call sets have much low sensitivity and relatively high FDRs, especially compared to SNPs

	Indels					SNPs	
	mpileup	**DINDEL**	**GATK**	**Union**		**GATK**	**Project VQSR***
No. of calls	38507	29730	97725	118316		516623	
All sites in CG 38							
True positives	11133	10531	21278	22758		276756	313969
False negatives	20506	21108	10361	8881		88712	51499
Hom-var sites in CG NA12878							
True positives	577	479	931	1055		24468	24031
False negatives	2025	2123	1671	1547		787	1224
Hom-var sites in GATK HiSeq NA12878							
True positives	3357	3115	4160	4546		25505	25124
False negatives	1371	1613	568	182		518	899
1000 Genomes Pilot 1 validation							
True positives	95	105	199	202		260	356
False positives	11	8	21	25		41	45
False negatives	160	150	56	52		157	61
Sensitivity	37.3	41.2	78.0	79.5		62.4	85.4
FDR (false discovery rate)	10.4	7.1	9.5	11.0		13.6	11.2

1000G Phase 1 EUR (Chr20); 351 samples, except for *Project VQSR over 1004 samples

Are we happy with the current calling results?

- Are these call sets sufficiently good to release as is?
- Do we need to devote additional resources to improve the methods?
  - May not have sensitivity we'd like, w.r.t. SNPs
  - Callers could tune up their sensitivity?
  - Do we want to explicitly genotype all indels in known data sets (Pilot 2, dbSNP)?
- Should we take the union of the calls?

Are we ready to carry out additional validation?

- Should we focus on using our comparative resources before additional validation?
  - CG 38 samples
  - Exomes
  - Comparisons to deep data sets?
- If we decide on validation:
  - Which indels should we validate?
  - What technology should we use? (Sequenom?)