BROAD INSTITUTE

1000 Genomes Phase I Exome indel validation: Results from new tech resequencing

Eric Banks, Yossi Farjoun

Genome Sequencing and Analysis
Medical and Population Genetics Program
Broad Institute of Harvard and MIT
October 16, 2012

Data and Definitions

- As part of the validation process for Phase 1 exome indel calls, we ran standard hybrid capture plus sequencing with novel techs on the 2 validation samples (NA10851 & NA19238) and NA12878.
    - HiSeq2000 with 104bp reads
    - MiSeq with 150bp reads
    - HiSeq2500 data not ready yet (run in the UK)

- Re-genotyping was performed with Broad's mapping-based caller using BCM's consensus Phase 1 calls as input alleles to confirm polymorphic status of sites
    - All discrepant or monomorphic sites were confirmed with manual inspection using IGV

Re-genotyping of consensus calls confirms polymorphic status of most sites

Phase 1 exome indel call set	No. calls in Project call set	No. confidently polymorphic	No. confidently monomorphic	No. where segregating allele is much larger event	No. unclear from current data *	Implied FDR of subset **
BCM's consensus calls* **	**297**	**273**	**12**	**3**	**9**	**5.2%**

* Unless calls from both techs were confidently concordant, status is "unclear"

** FDR = (confidently monomorphic + wrong allele) / (project calls – unclear)

*** Comparing against union of all center calls gave extremely poor sensitivity (~60%)

Some Caveats

- Note that this can't strictly be considered a validation of Project calls in its own right because the process is subject to the same error modes as the original data (capture/sequencing issues).
  - However, it is encouraging that the newer data and longer reads do confirm most of the original consensus calls.

- Also note that this analysis does not address the False Negative rate of the consensus call set.