2010/11/23 chr20 BAMs: data integrity analysis update & suggested samples for variant calling

Jan 4th, 2011

Hyun Min Kang & Goo Jun

University of Michigan

Recap & Overview

- Last call : Analyzed ILLUMINA+LS454 BAMs
  - 1,016 BAMs / 1,005 samples
  - Suggested samples to exclude :
    - 31 BAMs with genotype-free %MIX > 3%
    - 4 BAMs with very low library complexity
- Updates
  - Inclusion of 50 SOLiD BAMs
  - Per-sample depth
  - Initial suggestion of samples for variant calls
    - 1,008 samples across 1,031 BAMs

Summary of SOLiD data

- 50 sample BAMs
  - 15 BAMs with 25bp paired-end reads
  - 3 BAMs with 35bp single-end reads
  - 31 BAMs with 50bp paired-end reads
  - 1 BAM with 25bp + 50bp paired end reads
- 34 / 50 samples have data only in SOLiD platform
- Genotype data
  - 37 samples have HM3 genotypes
  - 36 samples have OMNI genotypes
  - 10 samples have neither HM3 nor OMNI genotypes

SOLiD 25bp reads

- Genotype-based % MIX
  - Ranges 3% ~ 9% (likely due to reference bias, not contamination)
- Genotype-free % MIX
  - Ranges 0% ~ 4%
- Reference bias measure 1 : refBase@HOMALT-Genotypes
  - Ranges 2% ~ 6%
- Reference bias measure 2 : Relative depth
  - [DP-HET]/[DP-REF] ranges 0.81-0.89
  - [DP-ALT]/[DP-REF] ranges 0.60-0.76
- Overall Evaluation
  - No obvious contamination
  - Slight concerns with reference bias
  - Suggestion : Include and evaluate genotyping accuracy

SOLiD 35-50bp reads

- Genotype-based % MIX
  - Ranges 0% ~ 2%
- Genotype-free % MIX
  - Ranges 0% ~ 1% for most sample
  - 1 Iberian sample (HG01630) show 17% of %MIX
- Reference bias measure 1 : refBase@HOMALT-Genotypes
  - Ranges 0.4% ~ 0.7%
- Reference bias measure 2 : Relative depth
  - [DP-HET]/[DP-REF] > 0.90
  - [DP-ALT]/[DP-REF] > 0.80
- Overall Evaluation
  - 1 suspected contamination
  - Other data look good
  - Suggestion : Exclude HG01630 and Include all the others

Combining all together

- Among 1,066 BAMs over 1,039 samples
  - 31 BAMs flagged for suspected contamination
  - 4 BAMs flagged for low library complexity
- Remaining samples to include :
  - 1,031 BAMs over 1,008 samples
- HapMap3 site depth distribution of the 1,008 samples
  - 4 samples < 1x *(Decided to exclude in today's call)*
    - NA18625(0.34x), NA20526(0.46x), HG01464(0.61x), NA19210(0.62x)
  - 26 samples < 2x
  - 127 samples < 3x
  - 13 samples > 10x
- **For chr20 calling : Use 1,004 samples / 1,027 BAMs**