Indel analysis: preliminary results

mtg-tech

01/11/2011
Lukas Habegger

Objective

- Analysis of indels at the nucleotide level
  - Manifestation of indels in the transcriptome:
    - Loss of function
      - Disruption of splice sites
      - Frameshifts
    - Trace indels: germline / somatic / transcriptome
    - Does it affect all transcripts or just a subset?
    - Effect on gene expression

Data

- Collaboration with Mark Rubin at Cornell

- 7 Matched prostate tissue samples
  - Normal (germline)
  - Tumor

- DNA sequencing done at the Broad
  - 30x coverage using Illumina

Dindel

- **Inputs**:
  - BAM file (read alignments)
  - FASTA file of the reference genome

- **Workflow**:
  - **Stage 1**: Extract *candidate indels* from BAM file
  - **Stage 2**: Group *candidate indels* into windows
  - **Stage 3**: Construct *candidate haplotypes* based on indels; realign reads to *candidate haplotypes* (computationally intense)
  - **Stage 4**: Indel calling and filtering; output in VCF format

Results

Number of indel calls (genome-wide) after PASS filtering

Sample	Normal	Tumor
STID0000000508	440660	476296
STID0000000581	412678	Error*
STID0000001701	488951	492736
STID0000001783	466196	457210
STID0000002832	475142	469113
STID0000003027	472286	463420
STID0000003043	481331	478858

* matepos inconsistency! (Reported from Dindel)

Number of indel calls (genome-wide) after PASS filtering
BAM files were preprocessed with SAMtools fixmate

STID0000000581	107428	39660
STID0000001783	68198	46294

Results continued

Raw number of indel calls (genome-wide)

	Normal				**Tumor**				**Totals**	
Sample	**Unique**		**Shared**		**Unique**		**Shared**		**Normal**	**Tumor**
STID0000000508	72372	11.1%	576940	88.9%	140433	19.6%	576940	80.4%	649312	717373
STID0000001701	100603	13.4%	650700	86.6%	108839	14.3%	650700	85.7%	751303	759539
STID0000001783	106501	15.6%	577751	84.4%	102006	15.0%	577751	85.0%	684252	679757
STID0000002832	100361	14.0%	615881	86.0%	95759	13.5%	615881	86.5%	716242	711640
STID0000003027	99236	14.0%	607274	86.0%	87012	12.5%	607274	87.5%	706510	694286
STID0000003043	87661	12.1%	638460	87.9%	77797	10.9%	638460	89.1%	726121	716257

Next steps

- Compare indels results to SNP calls
  - Problem: No SNP calls for germline
  - Potential solution: Use SNP arrays as a proxy

- Analyze the effect on gene expression
  - Map RNA-Seq reads to indel haplotypes