ChIP-Seq alignment pipeline proposal

Guiding principles:

- Utilize latest stable versions of software's that are currently in the production somewhere
  - To reduce running in unknown bugs at a consortium wide level

- Preserve data at the lowest threshold level currently applied among members
  - Don't throwaway data that someone can show will reliably yield signal

- Be explicit with parameters even when parameters are default
  - The command line should describe the algorithm up to implementation details instead of relying on manuals to get defaults
  - Command line should be in bam header

  IN: fastq, reference
  OUT: BAM suitable for enrichment analysis/peak calling

Status of current pipelines

- ENCODE (SET/PET)
  - BWA 0.7.7
    - [aln] seed length l = 32, mismatches in seed k = 2, dynamic read trimming q = 5
    - sex specific references for hg19
  - SAMtools 0.1.19
    - Not directly converted to BAM (bugs involving clipping of CIGAR strings not compatible with SAMtools)
    - Remove unmapped, non primary alignments, platform failed and duplicates -F1796 (1804)
    - SAMtools q = 30 filtering
  - Picard
    - Remove PCR duplicates with Mark Duplicates 1.110
- EBI (SET)
  - BWA 0.5.9
    - Dynamic read trimming q = 15
    - Gender matched references
  - SAMtools
    - [aln] q = 15 filtering
  - Picard
    - Mark duplicates
- DEEP
  - BWA 0.6.2 (convey machine architecture, cnybwa-0.6.2)
    - Dynamic read trimming [aln] q = 20
    - [sampe] -a 1000
  - SAMtools 0.1.19
    - Filtering -F 0x400 (no duplicates)

- CEMT
  - BWA 0.5.7
    - Chastity failed reads marked
    - [aln] all default parameters
    - [sampe] all default parameters
    - reference GRCh37lite http://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/
  - SAMtools 0.1.13
    - Filtering -F 1028 -q 5 (for analysis only, data submitted is unfiltered, PET fragments with length > 1000 are ignored)
  - Picard:
    - Mark Duplicates 1.71,

Notes:

- SAMtools 1.0 released
- OPTICAL_DUPLICATE_PIXEL_DISTANCE?

Straw man Proposal

BWA 0.7.7

[aln] seed length l = 32, mismatches in seed k = 2, dynamic read trimming q = 5

(make defaults explicit – can everyone send around a dump of BWA defaults, BAMPE vs BAMSE, how to deal with read length differences)

sex specific hg19 (20?)

RG tags + CL tags in bam header

Duplicates marked (Mark Duplicates 1.110)

No filtering for submitted data

Analysis:

SAMtools 0.1.19 (1.0+?)

q = 5 filtering, no chastity failed reads, no duplicates, (mate unmapped)

relevant SAMtools flags: 1804, 1796 (ENCODE), 0x0400 (DEEP), 1028 (CEMT)

Treatment of large fragments (> 1000bp, ignored at alignment level, analysis level, not at all?)