Epigenomics Characterization

Epigenomics

The USC Epigenome Center is currently involved in the mapping of the epigenetic state of human cells on a genome-wide scale. The Epigenomics workflow is essentially a data processing pipeline that uses the Pegasus Workflow Management System to automate the execution of the various genome sequencing operations. The DNA sequence data generated by the Illumina-Solexa Genetic Analyzer system is split into several chunks that can be operated on in parallel. The data in each chunk is converted into a file format that can be used by the Maq system. The rest of the operations involve the filtering out of noisy and contaminating sequences, mapping sequences into the correct location in a reference genome, generating a global map and then identifying the sequence density at each position in the genome This workflow is being used by the Epigenome Center in the processing of production DNA methylation and histone modification data.

Execution Profile

Execution times of Epigenomics jobs

Job	Count	Mean (s)	Variance
fast2bfq	146	0.39	0.02
fastqSplit	2	42	1.8e+02
filterContams	146	1.1	0.5
map	146	9635.01	1.7e+07
mapMerge	3	24	33
pileup	1	3269.73	0
sol2sanger	146	0.24	0.01

Sizes of Epigenomics data items

File Type	Count	Mean (MB)	Variance
chunked_sfq	420	7.3	0.18
filtered_sfq	420	5	0.096
fq_format	420	3.7	0.052
bfq_format	420	0.95	0.0045
out_map	420	1	0.0059
merged_map	6	68	18
merged_map	1	400.44	0
indexed_map	1	20	0
pileup	1	4.4	0