CONTAMINATOR - detect contaminating sequences in high-throughput sequencing data

Karl-Heinz Glatting, Güngör Budak, Lina Sieverling, and Agnes Hotz-Wagenblatt

Bioinformatics (Husar, W180), Core Facility Genomics&Proteomics, Deutsches Krebsforschungszentrum (DKFZ),
Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

home: http://genome.dkfz-heidelberg.de, genome@dkfz-heidelberg.de

Problem

One of the challenges in the analysis of data from next-generation instruments is sample contamination from non-sample sources like bacteria or viruses.

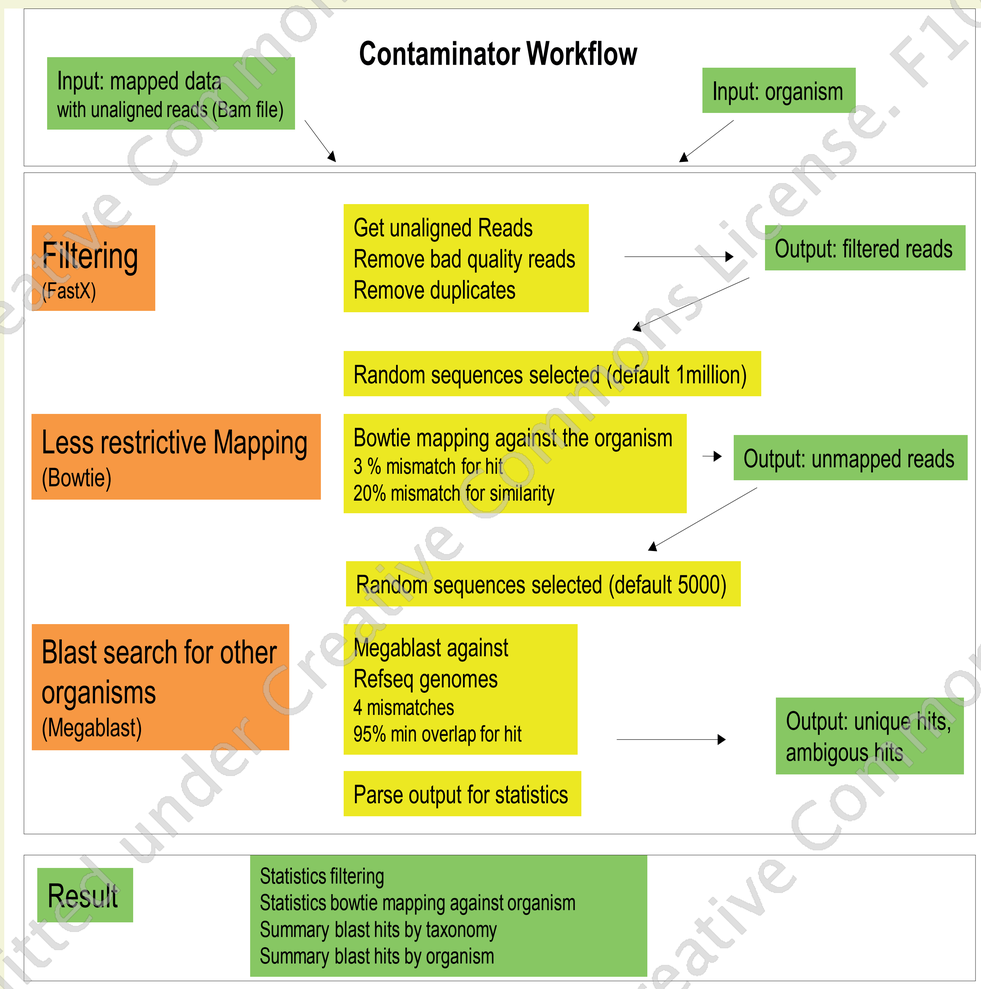
Solution

Contaminator is a new pipeline for the analysis of sequence reads that could not be mapped to the corresponding genome. It can be used to find out whether there was substantial contamination in your sequencing experiment and from which organism the reads come from.

Methods used

- Remapping a random subset of the unmapped reads (1,000,000 by default) with Bowtie against the corresponding genome, using less restrictive parameter settings.
- Blast search with randomly selected reads (5,000 by default) against genomic sequences from Refseq
- Hits are sorted by taxonomy and clustered into different taxonomic groups. The hits are classified either as "unique", if there is exactly one blast hit for a read fulfilling the threshold parameters, or as "ambiguous", if there are multiple hits with different organisms.

CONTAMINATOR Pipeline Design



Example SRR111919 (VACV infected HELA cells)

Beside testing samples which showed a large number of unmapped reads, we tested Contaminator with RNA-Seq data of virus infected cells (SRR111919), which should contain the vaccinia virus (VACV). The parameters were adapted to use more reads with Bowtie (5 000 000) and Megablast (1 000 000). The output with the identification of the virus is shown on the right side. Running time was approx. 24 h.

Conclusion

Contaminator can identify the origin of contaminations using a specified number of the reads that do not map. We could show that it works with substantial contamination e.g. with salmon DNA, but also with virus DNA in virus infected cells.
Contaminator has been implemented in the W3H task framework in the DKFZ (Ernst et al., Bioinformatics, 19:278-282; 2003).

Output

Contaminator

a tool for analysis of unmapped reads

Results for SRR111919.fastq.bam

Selected Parameters

Removal of duplicates	no
Minimum quality score of reads	0
Map unaligned reads with Bowtie	yes
Mapping genome for Bowtie	human_genome37
Number of random sequences to analyze with Bowtie	5000000
Maximum percent mismatch for read to be considered as hit	4
Maximum percent mismatch for read to be considered as similar to organism	20
Launch MegaBlast if at least this percentage of reads is unmapped	5
Taxonomic group(s) against which to map with MegaBlast	viruses
Number of random sequences to analyze with MegaBlast	1000000
Minimum overlap in percent for MegaBlast	95
Maximum number of mismatches for MegaBlast	4
Number of threads	5
Show unique hits by taxonomy starting at this number of hits	10
Show ambiguous hits by taxonomy starting at this number of hits	10
Show unique hits by organism starting at this number of hits	10
Show ambiguous hits by organism starting at this number of hits	10

Statistics filtering and mapping

Statistics

Filtering	Number
Reads in BAM file	22279096
Unaligned reads in BAM file	18256235
Unaligned reads that passed quality threshold	18256235

Bowtie	Number
Analyzed reads	5000000
Reads from organism	4078
Reads similar to organism	201298
Unmapped reads	4794624

MegaBlast	Number
Analyzed reads	1000000

MegaBlast Results

Unique Hits by Taxonomy

Name	Hits
Viruses, dsDNA viruses, no RNA stage, Poxviridae,	19102

Blast hits

Unique Hits by Taxonomy

Name	Hits
Viruses, dsDNA viruses, no RNA stage, Poxviridae, Chordopoxvirinae	19102
Viruses, dsDNA viruses, no RNA stage, Papillomaviridae, Alphapapillomavirus	206
Viruses, ssRNA negative-strand viruses, Bunyaviridae, Orthobunyavirus	60
Viruses, dsDNA viruses, no RNA stage, Herpesvirales, Alloherpesviridae	11

Ambiguous Hits by Taxonomy

Name	Hits
Viruses, dsDNA viruses, no RNA stage, Poxviridae, Chordopoxvirinae	60839

Unique Hits by Organism

Name	Hits
Vaccinia virus	18798
Alphapapillomavirus 7	206
Monkeypox virus Zaire-96-I-16	159
Taterapox virus	71
Simbu virus	51
Variola virus	28
Ectromelia virus	18
Cowpox virus	16
Camelpox virus	12
Anguillid herpesvirus 1	11

Ambiguous Hits by Organism

Name	Hits
Vaccinia virus	17031
Monkeypox virus Zaire-96-I-16	9552
Camelpox virus	9006
Ectromelia virus	8193
Taterapox virus	6645
Cowpox virus	5909
Variola virus	4502

download result in XML format

contaminator version 1.1
Created July 17, 2013 09:57 CEST

Vaccinia Virus hits shown!