From trash to treasure: detecting unexpected contamination in unmapped NGS data

Ilaria Granata, Mara Sangiovanni, Amarinder Singh Thind and Mario R. Guarracino
OUTLINE OF THE PRESENTATION

THE BIOLOGICAL FRAMEWORK

DECONTAMINER

CASE STUDIES
PART 1
THE BIOLOGICAL FRAMEWORK
ALIGNMENT TO THE REFERENCE GENOME

Adapted from: Meyerson, Gabriel & Getz. Nature Reviews Genetics. October 2010

Unmapped reads
WHY UNMAPPED READS?

• Low Quality Reads
• Repetitive Elements
• Sample/Reference differences
WHY UNMAPPED READS?

- Low Quality
- Repetitive Elements
- Sample/Reference differences
- Non-target organisms: CONTAMINATION

Unmapped reads
CONTAMINATION

Adapted from: Meyerson, Gabriel & Getz. *Nature Reviews Genetics*. October 2010}
HOW MANY UNMAPPED READS?

Breast Cancer RNA-seq data

Mapping stats:

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

<table>
<thead>
<tr>
<th>MAPQ</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPQ &gt;= 30</td>
<td>52939364.0</td>
<td>89.4</td>
</tr>
<tr>
<td>MAPQ &lt; 30</td>
<td>1256939.0</td>
<td>2.1</td>
</tr>
<tr>
<td>MAPQ &lt; 10</td>
<td>314858.0</td>
<td>0.5</td>
</tr>
<tr>
<td>MAPQ &lt; 3</td>
<td>4727781.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Unmapped</td>
<td>4727781.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Total</td>
<td>59238944.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.
How many unmapped reads?

Breast Cancer RNA-seq data

Mapping stats:

- MAPQ >= 30
- MAPQ < 30
- MAPQ < 20
- MAPQ < 10
- MAPQ < 3
- Unmapped

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPQ &gt;= 30</td>
<td>52939364.0</td>
<td>89.4</td>
</tr>
<tr>
<td>MAPQ &lt; 10</td>
<td>1256939.0</td>
<td>2.1</td>
</tr>
<tr>
<td>MAPQ &lt; 3</td>
<td>314558.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Unmapped</td>
<td>4727781.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Total</td>
<td>59238944.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Trash?
HOW MANY UNMAPPED READS?

Breast Cancer RNA-seq data

Mapping stats:

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

<table>
<thead>
<tr>
<th>MAPQ</th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPQ &gt;= 30</td>
<td>101996304.0</td>
<td>70.4</td>
</tr>
<tr>
<td>MAPQ &lt; 30</td>
<td>3509166.0</td>
<td>2.4</td>
</tr>
<tr>
<td>MAPQ &lt; 10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAPQ &lt; 10</td>
<td>882786.0</td>
<td>0.6</td>
</tr>
<tr>
<td>Unmapped</td>
<td>38501812.0</td>
<td>26.6</td>
</tr>
<tr>
<td>Total</td>
<td>144890054.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.
HOW MANY UNMAPPED READS?

Breast Cancer RNA-seq data

Mapping stats:

Number of alignments in various mapping quality (MAPQ) intervals and number of unmapped sequences.

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPQ &gt;= 30</td>
<td>10196304.0</td>
<td>70.4</td>
</tr>
<tr>
<td>MAPQ &lt; 30</td>
<td>3509166.0</td>
<td>2.4</td>
</tr>
<tr>
<td>MAPQ &lt; 10</td>
<td>882786.0</td>
<td>0.6</td>
</tr>
<tr>
<td>MAPQ &lt; 3</td>
<td>38501812.0</td>
<td>26.6</td>
</tr>
<tr>
<td>Unmapped</td>
<td>38501812.0</td>
<td>26.6</td>
</tr>
<tr>
<td>Total</td>
<td>144890054.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Lot of trash?
TRASH OR TREASURE?

UPSTREAM CONTAMINATION
APPLICATIONS TO PRECISION MEDICINE

GENOMICS
Our genes suggest what diseases we might be predisposed to.

PHENOTYPE
A complete status of health that can be used to prevent, diagnose & treat disease

LIFESTYLE/ENVIRONMENT
External factors such as diet, exercise & other lifestyle choices, medications, microbiota, and even where we live, influence our health.

MICROBIAL SIGNATURES

PROGNOSTIC MARKERS IN PRECISION MEDICINE
Overall non-vertebrates

Viruses

Bacteria

Plants

Large scale comparison of non-human sequences in human sequencing data

Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data
AIM:
  Detection of contaminant sequences

DETECTING UNEXPECTED CONTAMINATIONS

ANALYSIS OF UNMAPPED READS AS A STANDARD STEP AFTER ALIGNMENT

DECONTAMINER
PART 2
DECONTAMINER
DECONTAMINER: THE PIPELINE

Sequencing quality thresholds

- Match length, gap, mismatch thresholds
- Single/Paired end
- DNA-seq / RNA-seq
- Multiple samples
- BACTERIA
- FUNGI
- VIRUSES

Ambiguity resolution

Low quality | Ambiguous | Valid
DECONTAMINER: THE OUTPUT

BACTERIA

BLAST tables

Collected info

Summary

Fungi

Low quality matches

Ambiguous matches

Valid matches

VIRUSES

Match length or allowed gaps or mismatches below threshold

Paired reads not on the same species, or reads matching on different genera

Single sample MATCH COUNTS and stats

All samples MATCH COUNTS and stats

HTML reports
AIM:
Detection of contaminant sequences

BASIC FEATURES

• BASED ON A SUBTRACTION APPROACH
• RELIES ON SOFTWARES COMMONLY USED FOR THE ALIGNMENT PROCESS AND ON BASH/PERL SCRIPTS
• EASY TO INSTALL AND CONFIGURE IN A UNIX ENVIRONMENT
• ACCEPTS INPUT IN VARIOUS FORMATS (BAM, FASTA, FASTQ)
• RETURNS TEXTUAL OUTPUTS AND HTML-BASED REPORTS
• CURRENTLY DEVELOPED FOR HUMAN DATA VS BACTERIA, VIRUSES, FUNGI DATABASES
WHAT ABOUT OTHER SOFTWARE?

- **Taxonomer, SURPI, RNA-COMPASS, DeconSeq, PhymmBL**
  - Designed specifically for metagenomics, or based on 16S tags

- **PathSEQ**
  - Similar approach, but only available for Amazon Elastic Compute Cloud environment.

- **FastQ Screen, CaPSID**
  - Single sample, few contaminating organisms.

- **TruePure**
  - Analysis of only 10,000 random chosen reads.
Welcome to the download page of decontaMiner, a tool for detecting contaminating organisms in human unmapped sequences.

Next Generation Sequencing (NGS) experiments produce millions of short sequences that, mapped to a reference genome, provide biological insights at genomic, transcriptomic and epigenomic level.

Nonetheless, a variable number of reads fails to correctly align to the reference. In most of the cases this failure is due to the low quality of the bases called during the sequencing, but very often this ‘misalignment’ is due to sequence differences between the reads and the corresponding genome.

Investigating the source of these reads is definitely important to better assess the quality of the whole experiment, and to look for possible downstream or upstream ‘contamination’ from exogenous nucleic acids.

DecontaMiner is a tool designed and developed for the detection and analysis of these contaminating sequences. DecontaMiner is aimed at help researchers in obtaining more information from the data, and, in particular, to check for microorganisms presence that can not only affect the reliability of the whole experiment, but also foster the evaluation of the samples and the conditions under an additional perspective.

The download area below provides access to the following items:
- the current version of DecontaMiner code
- the installation and user guide
- the test data (please be patient, it is about 1.2 GB of data; the download may take some time)

An example of the interactive html report generated by DecontaMiner on the test data is available HERE.

Download area
- [decontaminer.tar.gz](http://www-labgtp.na.icar.cnr.it/decontaminer) Tar gzipped archive of the decontaminer software
- [InstallationAndUserGuide.pdf](http://www-labgtp.na.icar.cnr.it/decontaminer) DecontaMiner installation and user guide
- [test_data.tar.gz](http://www-labgtp.na.icar.cnr.it/decontaminer) Tar gzipped archive of the single-end unmapped bam test data
PART 3
CASE STUDIES
**RNA-Seq DATA**

**PRE-PROCESSING PIPELINE**
- FASTQC
- FASTX (Trimmer)
- TOPHAT2 (hg19; UCSC .gtf)

**GEO DATASET (GSE68086)**

“RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass and molecular pathway cancer diagnostics”
- 228 tumor-educated platelet (TEP) (non-small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer and hepatobiliary carcinomas)
- 55 healthy samples

**UNPUBLISHED DATASET**

Different Breast Cancer subtypes, healthy tissues form cancer patients, and healthy controls
- 16 tumoral (Lum-A, Lum-B, TNBC, Her2)
- 4 Healthy tissues
- 4 Healthy controls
BLOOD PLATELETS DATASET

BASIC STATISTICS FOR BACTERIA CONTAMINATION - GROUPING BY SPECIES

Sample information

- Number of sample with contamination: 230
- Number of sample with contamination above MC threshold: 150

Filtering Parameters

- Match Length: 101
- GAP Number: 0
- Mismatch Number: 0
- Match count (MC) threshold: 100

Organism name: Propionibacterium acnes, sample id: Blood_Platelets_GBM-409_umm, contamination: 96.8%
INFORMATION FOR SAMPLE Blood_Platelets_CRC-272_unm

BASIC STATISTICS

<table>
<thead>
<tr>
<th>Number of Alignments</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>17805423</td>
</tr>
<tr>
<td>Low-quality</td>
<td>12959964</td>
</tr>
<tr>
<td>Hi-quality</td>
<td>48845459</td>
</tr>
<tr>
<td>Ambiguous</td>
<td>308179</td>
</tr>
<tr>
<td>Valid</td>
<td>4537280</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Reads</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>839307</td>
</tr>
<tr>
<td>Hi-quality</td>
<td>639540</td>
</tr>
<tr>
<td>Valid</td>
<td>635938</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Organisms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>3826</td>
</tr>
<tr>
<td>Hi-quality</td>
<td>2005</td>
</tr>
<tr>
<td>Valid</td>
<td>1001</td>
</tr>
<tr>
<td>Valid above the MC</td>
<td>39</td>
</tr>
</tbody>
</table>
**Propionibacterium acnes**: Disease-Causing Agent or Common Contaminant? Detection in Diverse Patient Samples by Next-Generation Sequencing

Sarah Mollerup, Jens Friis-Nielsen, Lasse Vinner, Thomas Arn Hansen, Stine Raith Richter, Helena Fridholm, Jose Alejandro Romero Herrera, Ole Lund, Søren Brunak, Jose M. G. Izarzugaza, Tobias Mourier, Lars Peter Nielsen, Anders Johannes Hansen
BREAST CANCER DATASET: BACTERIAL CONTAMINATION

The diagram illustrates the percentage of various bacterial species found in different breast cancer subtypes. The subtypes are categorized as Healthy controls, Healthy tissues, Her-2, Luminal A, Luminal B, and TNBC.

- **Acinetobacter baumannii** is highlighted in red.
- **Bacillus coagulans** is shown in dark blue.
- **Bacillus subtilis** is represented in light blue.
- **Escherichia coli** is indicated in green.
- **Geobacillus sp.** is marked in purple.
- **Propionibacterium acnes** is depicted in teal.
- **Pseudomonas aeruginosa** is shown in light green.
- **Pseudomonas fluorescens** is represented in dark green.
- **Pseudomonas putida** is in light blue.
- **Staphylococcus aureus** is indicated in dark blue.
- **Staphylococcus epidermidis** is shown in purple.
- **Streptococcus agalactiae** is depicted in teal.
- **Streptococcus pneumoniae** is marked in light green.
- **Streptococcus pyogenes** is represented in dark green.
- **Streptococcus suis** is in light blue.
- **Others** are indicated in yellow.

The data shows a significant variation in bacterial contamination across different subtypes, with some subtypes showing higher percentages of certain bacteria compared to others.
BREAST CANCER DATASET: BACTERIAL CONTAMINATION
Microbiota of Human Breast Tissue

Camilla Urbania,⁎,⁎, Joanne Cummins,⁎, Muriel Brackstone,⁎, Jean M. Macklain,⁎,⁎, Gregory B. Gloor,⁎, Chwanrow K. Baban,⁎
Leslie Scott,⁎, Deirdre M. O’Hanlon,⁎, Jeremy P. Burton,⁎, Kevin P. Francis,⁎, Mark Tangney,⁎, Gregor Reid⁎,⁎

Lawson Health Research Institute, London, Ontario, Canada; Department of Microbiology and Immunology, Western University, London, Ontario, Canada; Cork Cancer Research Centre, University College Cork, Cork, Ireland; London Regional Cancer Program, London, Ontario, Canada; Department of Biochemistry, Western University, London, Ontario, Canada; Department of Surgery, Cork University Hospital, Cork, Ireland; Preclinical Imaging, Perkin Elmer, Alameda, California, USA

In recent years, a greater appreciation for the microbes inhabiting human body sites has emerged. In the female mammary gland, milk has been shown to contain bacterial species, ostensibly reaching the ducts from the skin. We decided to investigate whether there is a microbiome within the mammary tissue. Using 16S rRNA sequencing and culture, we analyzed breast tissue from 81 women with and without cancer in Canada and Ireland. A diverse population of bacteria was detected within tissue collected from sites all around the breast in women aged 18 to 90, not all of whom had a history of lactation. The principal phylum was Proteobacteria. The most abundant taxa in the Canadian samples were Bacillus (11.4%), Acinetobacter (10.0%), Enterobacteriaceae (8.3%), Pseudomonas (6.5%), Staphylococcus (6.5%), Propionibacterium (5.8%), Comamonadaceae (5.7%), Gammaproteobacteria (5.0%), and Prevotella (5.0%). In the Irish samples the most abundant taxa were Enterobacteriaceae (30.8%), Staphylococcus (12.7%), Listeria welshimeri (12.1%), Propionibacterium (10.1%), and Pseudomonas (5.3%). None of the subjects had signs or symptoms of infection, but the presence of viable bacteria was confirmed in some samples by culture. The extent to which these organisms play a role in health or disease remains to be determined.

May 2014
CONCLUSIONS

• **UNMAPPED READS** MIGHT BE THE SOURCE OF PRECIOUS INFORMATION

• **DECONTAMINER** IS A TOOL TO INVESTIGATE THE PRESENCE OF UPSTREAM/DOWNSTREAM CONTAMINATION IN HUMAN NGS DATA

• **DECONTAMINER** IS FREE, EASY TO INSTALL AND USE
FUTURE WORK

• PLANTS AS A REFERENCE AND A CONTAMINATING DATABASE

• COMPANION WEB-SITE WITH MACHINE LEARNING TOOLS

• DEVELOPMENT AS A DOCKER CONTAINER
  (https://www.docker.com/what-docker)

• INTEGRATION IN GALAXY
  (https://www.galaxyproject.org)
THANKS TO

Dr. Mario R. Guarracino

Dr. Ilaria Granata

Amarinder Singh Thind

http://www-labgtp.na.icar.cnr.it/decontaminer

“The real voyage of discovery consists not in seeking new landscapes, but in having new eyes.”

Marcel Proust
• Paired Ends
• FastQ quality
  • Minimum quality score to keep: 20
  • Minimum percent of bases that must have ≥20 quality: 100%
• SortMeRNA
  • E-value ≤ 10^{-20}
• BLAST
  • Match length = Query length = 101 nt
  • Mismatches N°: 0
  • Gaps N°: 0