Comprehensive analysis of RNA-sequencing to find the source of every last read across 544 individuals from 53 tissues

Serghei Mangul
Eskin Lab (ZarLab)
University of California, Los Angeles
sergheimangul.wordpress.com

@serghei_mangul
RNA-Seq

Assemble novel transcripts

(Differential) gene expressions

RNA splicing
RNA-Seq reads

In-silico separation

Human references

Mapped human reads

Unmapped reads

Read Origin Protocol

Genotype-Tissue Expression (GTEx v6)

- 8,555 samples
  - across 544 individuals
  - from 53 body sites

55.4M paired-end reads

Map to the human genome (tophat2)

Mapped reads (87.2%)
Unmapped reads (12.8%)

Genomic profile of mapped reads (gprofile)

Genomic profile of repeat elements (rprofile)

Read Origin Protocol
1. Quality control (FASTQC, SEQCLEAN)

2. Remap to human references (Megblast)

RNA-Seq

55.4M paired-end reads

- Low quality reads (4.8%)
- rRNA repeat (2.0%)
- Lost human reads (4.8%)
55.4M paired-end reads

RNA-Seq

Genomic profile of unmapped reads

a. Quality control (FASTQC, SEQCLEAN)

b. Remap to human references (Megablast)

c. Map to repeat sequences (Megablast)

- low quality reads (4.8%)
- rRNA repeat (2.0%)
- lost human reads (4.8%)
- lost repeat elements (0.03%)
Repeat profile

*On average 7% (~3M) reads are categorized as repeats.
RNA-Seq

55.4M paired-end reads

a. Quality control (FASTQC, SEQCLEAN)

b. Remap to human references (Megblast)

c. Map to repeat sequences (Megablast)

d. Non-co-linear (NCL) RNA profiling (ncSplice)

Genomic profile of unmapped reads

low quality reads (4.8%)

rRNA repeat (2.0%)

lost human reads (4.8%)

lost repeat elements (0.03%)

NCL RNA (0.1%)

Trans-splicing (gene fusion)

Read Origin Protocol
Adaptive immune repertoires

mRNA encoding for T and B cell receptor chains
Quality control (FASTQC, SEQCLEAN)

Remap to human references (Megablast)

Map to repeat sequences (Megablast)

Non-co-linear (NCL) RNA profiling (ncSplice)

B and T cell receptors profiling (ImReP)

55.4M paired-end reads

RNA-Seq

Mapped reads

low quality reads (4.8%)

rRNA repeat (2.0%)

lost human reads (4.8%)

lost repeat elements (0.03%)

NCL RNA (0.1%)

immune reads (0.02%)
Adaptive immune repertoires

*IGH: immunoglobulin heavy chain
ROP(ImRep) is able to identify samples with high activity of lymphocytes

http://www.gtexportal.org/home/histologyPage#data
55.4M paired-end reads

RNA-Seq

- **a. Quality control (FASTQC, SEQCLEAN)**
- **b. Remap to human references (Megablast)**
- **c. Map to repeat sequences (Megablast)**
- **d. Non-co-linear (NCL) RNA profiling (ncSplice)**
- **e. B and T lymphocytes profiling (ImReP)**
- **f. Microbiome profiling (MCS/Megablast)**

- **Genomic profile of unmapped reads**
  - low quality reads (4.8%)
  - rRNA repeat (2.0%)
  - lost human reads (4.8%)
  - lost repeat elements (0.03%)
  - NCL RNA (0.1%)
  - V(D)J recombinations (0.02%)
  - microbial reads (0.2%)
  - unaccounted reads (0.9%)

- **Read Origin Protocol**

  - template jumping
  - unknown microbes (EMDeBruijn*)
  - hyper-editing (E. Levanon)

Human microbiome

a. Superkindom

b. Phyla

- Proteobacteria: 67%
- Actinobacteria: 24%
- Firmicutes: 8%

i. Betaproteobacteria: 53%
ii. Gammaproteobacteria: 16%
iii. Actinobacteria: 24%

iv. Bacilli: 5%

v. Burkholderiales: 11%
vi. Pseudomonadales: 53%

vii. Enterobacteriales: 67%

viii. Actinomycetales: 24%
ix. Bifidobacteriales: 8%

x. Lactobacillales: 11%

Positive Control: Cells - EBV-transformed lymphocytes samples

Obtained by MetaPhlAn2

Microbial Coverage Scanner (MCS v1.0)
ROP tool

No installation is required. The ROP comes with no pre-requirements except Python 2.7.

ROP Tutorial

- What is ROP?
- How ROP works?
- How to install ROP?
- Get started
- ROP analysis: one RNA-Seq sample
- ROP output details
- How to map reads and save unmapped reads?
- Source of every last read

https://github.com/smangul/rop/wiki

ROP is available at https://sergheimangul.wordpress.com/rop/

Typical output of ROP

- RNA-Seq
- WGS
- WES
- Chip-Seq
- Single cell Seq

Read Origin Protocol

- Lost human reads
- Lost repeat elements
- Non-co-linear RNAs
- V(D)J recombination (BCR/TCR)
- Microbial communities
Acknowledgment

Harry Taegyun Yang
Maura Rossetti
Roberto Spreafico
Loes Loohuis
Roel A. Ophoff
Eleazar Eskin

Noah Zaitlen
Mark Ansel
Nicolas Strauli
Stephanie Christenson
Ryan D. Hernandez
Prescott G. Woodruff

B.I.G. Summer Institute
SUMMER RESEARCH EXPERIENCE IN GENOMICS AND BIOMEDICINE

Kevin Wesel
Will Van Der Wey
Jeremy Rotman
Benjamin Statz

Sagiv Shifman

GTEx Consortium

David Koslicki

Franziska Gruhl
ROP protocol can find lost human reads

reads within the
threshold of
tophat2

reads with additional mismatches and/or short
gaps

Percentage of lost human reads (%)
Microbiome diversity is significantly different across tissues

Alpha diversity

Adrenal gland  Heart  LCL  Lung  Hypophysis  Thyroid

Negative Controls

p<10^-16

Obtained by MetaPhlAn2