updated on 2021-04-14

Single cell cDNA mutation calling:

performance comparison between Loop Genomics and R2C2

#### Objective

* We will select 2 cell lines with distinct mutation profiles, and mix them in an experiment to evaluate the specificity of Loop Genomics' method.
  + NCI-H358 and K-562 (blue columns in the table below), both of which I have been using, could be two convenient options. They are derived from lung and blood respectively, so we should be able to easily distinguish them in the accompanying single-cell RNAseq data.
  + Mutations in the table below are based on either Depmap <https://depmap.org> or COSMIC <https://cancer.sanger.ac.uk/cell_lines>. The unverified or conflicting information is marked as "maybe".
* We will select a panel of genes (20) to cover a wide range of expression levels and transcript lengths.
  + In the table below, L= mRNA length.
  + In the table below, numbers indicate expression (nRPKM) for each gene in each cell line (data source: <http://resdev.gene.com/genehub6/#/summary/gene/ENSG00000120693/4093/human>)

#### Table 1, Potential genes of interest for cDNA enrichment and mutation calling

|  |  | **Lung** | | **Blood** | **Colon** | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **HCC1171** | **NCI-H358** | **K-562** | **LS-180** | **SW403** |
| 1 | DNAJB1  L=2233 | 49 | 46 | 175  S186= (c.558C>A) | 317 | 71 |
| 2 | CTNNB1  L=3720 | 89.4 | 142.7  T75A | 43.1 | 84.3  S45F (homo) | 87.6 |
| 3 | KRAS  L=5436 | 81.8  G12C | 22.6  maybe G12C | 7.3 | 7.4  G12D | 8.1 maybe G12V |
| 4 | TP53  L=2591 | 18.7  N247I (homo) | 0.01  homozygous deletion? | 1.8  maybe Q136fs | 49.7 | 9.4 |
| 5 | BCL6  L=3575 | 2.9 | 4.3  A45S | 10.7 | 3.6 | 1.1 |
| 6 | CBL  L=11241 | 9 | 7.7 | 14  D792H | 3.5 | 3.9 |
| 7 | FANCC  (L=4612) | 3.7 | 5  S57C | 3 | 1.7  maybe intronic | 3.7 |
| 8 | JAK2  L=5285 | 2.0 | 2.2  R1117M | 14 | 3.1 | 1.5 |
| 9 | TP63  L=4927 | 0.5 | 6.5  maybe V364I | 0.36 | 0.03 | 0.008 |
| 10 | ERBB3  L=5765 | 0.14 | 8.8 | 2.7  R611W | 15 | 49.9 |
| 11 | DROSHA  L=5424 | 32 | 29  maybe R219I | 17 | 5.8  maybe R1066H | 16 |
| 12 | ASXL1  L=7056 | 18 | 14.6 | 20  Y591\* | 6 | 11  maybe S444\* |
| 13 | EGFR  L=5616 | 17.7 | 10.8 | 0.002 | 6.8 | 3.1 |
| 14 | HOXA9  L=2076 | 0.07 | 0.001 | 0.001  D192N | 30.4 | 17.5 |
| 15 | HOXD11  L=1463 | 2.65 | 0.001 | 0.001  E294K | 0.001 | 0.001 |
| 16 | SMAD9  L=5592 | 1.5 | 0.07  maybe H130N | 0.02 | 1.6 | 0.1 |
| 17 | FGFR1  L=5375 | 4 | 1  maybe R240H | 0.06 | 0.04 | 0.2 |
| 18 | PTPRC  L=5429 | 0.001 | 0.01  M769K | 21 | 0.01 | 0.06 |
| 19 | BRAF  L=2949 | 4.6 | 3.7 | 15.2 | 3.2  D211G | 0.8 |
| 20 | APC  L=10740 | 1.9 | 3.8 | 3.8 | 4.7  R1788C | 3.8 |
| 21 | PI3KCA  L=3724 | 2.3 | 2.9 | 5.5 | 1.8  H1047R | 2.6  maybe Q546K |
| 22 | PTEN  L=5572 |  |  |  |  |  |
| 23 | ALK  L=6267 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

#### gCell RNAseq data analysis results

##### In-house RNAseq data location:

NGS ID 171: SAM635643 NCI-H358 LIB2525 POOL1647 and POOL1179

fastq files: /gne/data/dnaseq/processed\_runs/R250/results

LIB2525\_SAM635643\_L7\_R1.fastq.gz

LIB2525\_SAM635643\_L7\_R2.fastq.gz

NGS ID 171: SAM636855 K-562 LIB6663

fastq files: /gne/data/dnaseq/processed\_runs/R1048/results

LIB6663\_SAM636855\_L4\_R1.fastq.gz

LIB6663\_SAM636855\_L4\_R2.fastq.gz

NGS1799 data analysis for gCell RNA Seq access denied (this may not be very useful)

This is optional: NGS2535 gCell cell lines stranded RNA-seq: need to know which SAM ID is which cell line

##### Analysis procedure:

1, Using STAR to align (done)

Alignment results:

Rosalind: /gne/data/dnaseq/analysis/wub31/gCellMutation/NCIH358/STARoutput/

Rosalind: /gne/data/dnaseq/analysis/wub31/gCellMutation/K562/STARoutput/

2, Sort and index (done)

Results are stored in the same STARouput folders

3, example: reads pileup at KRAS G12C for NCI-H358(done)

[mpileup] 1 samples in 1 input files

12 25245351 C 306 <<<<<<<<<<,$,aaaa,aaa.AAAAa,aaaaaaaA.aaaaaa,aaa,,Aaa,aAAaa,aaAA.Aa.aa,aa,,aa,a..A.AAAAAA,aaAAAAaaaaa,aaaaaAaaaAAA.,aaaa.aaAa,a,aa,a,aaaaaaaaa,a,,aaaaaa,,aaaaaaaaaaa,a,a,,aaaaaa,aa,aaaaaaaaaa,aA.AAa.aaa,aaa,aaaaaaa,,aaaaa,,a,aaaaaaaa,aaaaaaAaaaaaa,aa,,aaaa,aa,,,a,,aaa,.aaa,aaA.a,aaAAAaa.A.AaAa,a,aa,aaa..^~,^~,^~, II@FEB2???B@====C===o>qHkFFFFFFDFFHHHHHHHHHHHHHHJHJJJrJHIJIJJsoJJIJJGJJJJFEIJJJlgIEJJJJJJJIJGJJIJGEJJJJCJJJI?IJJEJJJJGJJFJJJJIJG1JJI@IJJGJJJJJJJJIJGHHIIJJGJJJJHIIJIGDJJJIJJJIJJJJJJJGFJJJGJIJFHJJJJJJJBJJIFGJJIIJJHJJJIJIIIJJJJJJJIIJJIIJJJJJIHHAFFDFFFFFFFFFFFFFDFCFF?FFCF@CE>JEDFD@DDF2FD;FFDFCCDDDDDDADDCCFDCC

Conclusion: This confirms that most reads have the C-->A mutation as expected.

4, Estimate zygosity (I am using % of mutated reads in all reads spanning the position as an approximation)

5, Summary for all mutations of interest (done)

##### Table 2, confirmed mutations and their prevalence

| gene | mutation (amino acid) | position on GRCh38 | NCI-H358 RNAseq data | K-562 RNAseq data |
| --- | --- | --- | --- | --- |
| DNAJB1 | S186= | 19:14516700-14516700 | no mutation | G-->T 33% |
| CTNNB1 | T75A | 3:41224735-41224735 | A-->G 33% | no mutation |
| KRAS | G12C | 12:25245351-25245351 | C-->A 72% | no mutation |
| TP53 | Q136fs | 17:7675205-7675206 | deletion 100% | insertion of a G 5%,20% |
| BCL6 | A45S | 3:187733561-187733561 | C-->A 28% | no mutation |
| CBL | D792H | 11:119298480-119298480 | no mutation | G-->C 33% |
| FANCC | S57C | 9:95247512-95247512 | G-->C 26% | no mutation |
| JAK2 | R1117M | 9:5126742-5126742 | G-->T 54% | no mutation |
| TP63 | V364I | 3:189868677-189868677 | G-->A 49% | no mutation |
| ERBB3 | R611W | 12:56094528-56094528 | no mutation | C-->T 79% |
| DROSHA | R219I | 5:31526277-31526277 | C-->A 28% | C-->G 0.6% (1/159 reads) |
| ASXL1 | Y591\* | 20:32434485-32434485 | no mutation | C-->A 29% |
| EGFR | none reported in NCI-H358 and K-562 |  |  |  |
| HOXA9 | D192N | 7:27164884-27164884 | no reads | no reads |
| HOXD11 | E294K | 2:176109005-176109005 | no reads | no reads |
| SMAD9 | H130N | 13:36879301-36879301 | no reads | no reads |
| FGFR1 | R240H | 8:38426241-38426241 | C-->T 73% | no reads |
| PTPRC | M769K | 1:198735155-198735155 | T-->A (1 read coverage) | No mutation (335 reads) |
| BRAF | none reported in NCI-H358 and K-562 |  |  |  |
| APC | none reported in NCI-H358 and K-562 |  |  |  |
| PI3KCA | none reported in NCI-H358 and K-562 |  |  |  |
| PTEN | none reported in NCI-H358 and K-562 |  |  |  |
| ALK | none reported in NCI-H358 and K-562 |  |  |  |

Questions : answers updated on April 26

Do you have any data to show ? How do we really know it works ?

How much material ? → For single-cell experiments, they recommend doing the pcr amplification in a real time fashion and stop the amplification before the plateau stage which helps to reduce chimera.

How can we directly compare to long reads post selection ?

Can we get the oligo design to order the capture probes ourselves ? → They will use either IDT or Twist probes depending on scale. They will send us the oligo design.

What info do you need from us other than gene names ? → gene names or fasta files

Is the price still 10K for 4 samples ?

Timeline?

We will also test a synthetic transcript mix (Lexogen SIRV-set 4) (Is this mix compatible, transcript lengths, number, complexity etc.) RT conditions → They take the RNA directly, that makes this easy, send a few ul of few ng/ul.

Table 3. List of 50 genes of interest

KRAS

BRAF

CTNNB1

ERBB3

FGFR1

MAP2K1

NRAS

PIK3CA

RAF1

STK11

KEAP1

APC

TP53

AKT1

ALG3

ARAF

CBL

CRK

DUSP14

DUSP7

EGFR

ERBB2

FGFR2

FGFR3

FGFR4

HRAS

MAP2K7

MAP3K11

MAP3K12

MAP4K1

MAPK1

MAPK14

MAPK3

MAPK7

MAPK9

MET

MYC

NTRK2

PIK3R1

PTCH1

PTCH2

ROR1

RPS6KA1

RPS6KA5

RRAS

SHOC2

SOS1

SOS2

SRC

YWHAE