Reading PoolQ Files, Assessing Screen Quality, and Hit Call Considerations

Guidance from the GPP regarding analysis

following a standard pooled screening experiment

Updated 10/15/2020

This is a frequently updated file, please check back regularly for the current version

at the Pooled Screen Analysis Guide link at
https://portals.broadinstitute.org/gpp/public/

For reference and more in depth discussion see literature published from the GPP regarding general pooled-screening design and analysis:

Piccioni F, Younger ST, Root DE, *Curr Protoc Mol Biol*, 2018 Jan 16;121:32.1.1-32.1.21

Doench JG, *Nat Rev Genet*, 2018 Feb;19(2):67-80

Doench JG, Hanna RE, *Nat Biotechnol*, 2020 Jul;38(7):813-823

Overview

- What is in the PoolQ analysis you received from GPP?
- Graphs you should look at to determine the quality of your data
- Calling your hits: some guidelines

What is in the PoolQ analysis you received from GPP?

First, we'll guide you through the PoolQ analysis provided by the GPP

When you click the link sent to you in your email you should see the following menu of links and files:

PoolQ Analysis: Example_Screen

Sample
└→ Example_Screen
Clone Pool
└→ CP0041
Run Date
└→ 2020-10-08 17:30

PoolQ Input Files

	Download	Filename	Description
1	.csv	Example_Screen_Conditions.csv	CSV file mapping sample barcodes to experimental conditions
2	.csv	CP0041_reference_20160112.csv	CSV file mapping construct barcodes to barcode identifiers
3	.gz	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes
4	.gz	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes

PoolQ Output Files

	Download	Filename	Description
1	.txt	counts-Example_Screen.txt	PoolQ counts matrix
2	.txt	quality-Example_Screen.txt	PoolQ quality report
3	.txt	lognorm-Example_Screen.txt	Log-normalized counts matrix
4	.txt	correlation-Example_Screen.txt	Condition scores correlation matrix
5	.txt	barcodecounts-Example_Screen.txt	Counts by condition barcode
6	.txt	runinfo-Example_Screen.txt	PoolQ runtime information
7	.txt	data-integrity-Example_Screen.txt	Janssen discovery data integrity (DDI) document [?]

FASTQC Reports

	Link	Download	Filename	Description
1	.html	.zip	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes FASTQC analysis
2	.html	.zip	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes FASTQC analysis

→ Go to the pooled screen analysis page

Now, let's look at some of the PoolQ outputs: First, we highlight the CHIP file link.

PoolQ Analysis: Example_Screen

Sample
└→ Example_Screen
Clone Pool
└→ CP0041 ← CHIP file link
Run Date
└→ 2020-10-08 17:30

PoolQ Input Files

	Download	Filename	Description
1	.csv	Example_Screen_Conditions.csv	CSV file mapping sample barcodes to experimental conditions
2	.csv	CP0041_reference_20160112.csv	CSV file mapping construct barcodes to barcode identifiers
3	.gz	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes
4	.gz	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes

PoolQ Output Files

	Download	Filename	Description
1	.txt	counts-Example_Screen.txt	PoolQ counts matrix
2	.txt	quality-Example_Screen.txt	PoolQ quality report
3	.txt	lognorm-Example_Screen.txt	Log-normalized counts matrix
4	.txt	correlation-Example_Screen.txt	Condition scores correlation matrix
5	.txt	barcodecounts-Example_Screen.txt	Counts by condition barcode
6	.txt	runinfo-Example_Screen.txt	PoolQ runtime information
7	.txt	data-integrity-Example_Screen.txt	Janssen discovery data integrity (DDI) document [?]

FASTQC Reports

	Link	Download	Filename	Description
1	.html	.zip	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes FASTQC analysis
2	.html	.zip	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes FASTQC analysis

→ Go to the pooled screen analysis page

What are CHIP files?

- CHIP files map barcodes (eg. sgRNAs) to genes
- CHIP files are updated regularly to reflect changes to genome annotations
- Multiple versions are provided, use the one marked “Preferred" (uses of the other alternative versions can be explained as needed)
- All screens being compared to one another should use the same CHIP file
- CHIP files are tsv formatted text files with three columns: 1. Barcode (eg. sgRNA), 2. Gene Symbol, and 3. Gene ID for any matching genes. *Note: Barcodes that match more than 1 gene will be listed in multiple rows.*
- CHIP files can be opened in EXCEL and Text Edit, as well as other applications

The following table should appear after you have clicked on the CHIP file link from your PoolQ output page. Here the preferred CHIP file version is boxed in red.

	Type	File	Last Changed	Target	Match	Description
1	Reference [?]	CP0041_reference_20160112.csv	2016-01-12	n/a	n/a	Lists all 20mer barcodes contained in the pool, with their associated construct ID(s), if any. Used during initial sequence deconvolution, e.g. with PoolQ.
2	CHIP [?]	CP0041_17mer_20180831.chip	2018-08-31	gene	lax [?]	(LEGACY) 20mer barcode to gene mapping via lax [?] match to CDS regions only.
3	CHIP [?]	CP0041_GRCh38_NCBI_lax_gene_20200612.chip	2020-06-12	gene	lax [?]	20mer barcode to gene mapping via lax [?] sgRNA sequence match to NCBI-annotated genes in GRCh38 primary assembly.
4	CHIP [?]	CP0041_GRCh38_NCBI_strict_gene_20200612.chip	2020-06-12	gene	strict [?]	**(PREFERRED)** 20mer barcode to gene mapping via strict [?] sgRNA sequence match to NCBI-annotated genes in GRCh38 primary assembly.
5	CHIP [?]	CP0041_origtarget_20191021.chip	2019-10-21	gene	n/a	20mer barcode to gene mapping using construct's originally intended target gene, if known.
6	BED [?]	CP0041_9606_GRCh37_20181031_ontarget.bed	2018-10-31	n/a	n/a	Genome annotation track data in the UCSC BED format, with all perfect sgRNA target sequence (20mer) matches at PAM sites in target assembly GRCh37.
7	BED [?]	CP0041_9606_GRCh38_20181031_ontarget.bed	2018-10-31	n/a	n/a	Genome annotation track data in the UCSC BED format, with all perfect sgRNA target sequence (20mer) matches at PAM sites in target assembly GRCh38.

Download CSV

There are some projects involving unique custom clone pools with unsupported elements for which a CHIP file doesn't exist. Bear with us, and we will work with you to create the custom files you may require for your analysis.

Inside the PoolQ Link provided to you by GPP: Quality File

PoolQ Analysis: Example_Screen

Sample
└→ Example_Screen
Clone Pool
└→ CP0041
Run Date
└→ 2020-10-08 17:30

PoolQ Input Files

	Download	Filename	Description
1	.csv	Example_Screen_Conditions.csv	CSV file mapping sample barcodes to experimental conditions
2	.csv	CP0041_reference_20160112.csv	CSV file mapping construct barcodes to barcode identifiers
3	.gz	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes
4	.gz	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes

PoolQ Output Files

	Download	Filename	Description
1	.txt	counts-Example_Screen.txt	PoolQ counts matrix
2	.txt	quality-Example_Screen.txt	PoolQ quality report
3	.txt	lognorm-Example_Screen.txt	Log-normalized counts matrix
4	.txt	correlation-Example_Screen.txt	Condition scores correlation matrix
5	.txt	barcodecounts-Example_Screen.txt	Counts by condition barcode
6	.txt	runinfo-Example_Screen.txt	PoolQ runtime information
7	.txt	data-integrity-Example_Screen.txt	Janssen discovery data integrity (DDI) document [?]

Quality File

FASTQC Reports

	Link	Download	Filename	Description
1	.html	.zip	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes FASTQC analysis
2	.html	.zip	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes FASTQC analysis

→ Go to the pooled screen analysis page

Quality file from PoolQ

Total reads: 171425684
Matching reads: 134214114
1-base mismatch reads: 0

Overall % match: 78.29

Reads with no construct barcode: 7775868
Max construct barcode index: 30
Min construct barcode index: 11
Avg construct barcode index: 25.68

Read counts for sample barcodes with associated conditions:

Barcode	Condition	Matched (Construct+ Sample Barcode)	Matched Sample Barcode	% Match	Normalized Match
TTGAACCG	Sample 1_Drug_Rep 1	1753194	2045733	85.7	13.32
AATCCAGC	Sample 1_Drug_Rep 1	1811679	2108036	85.94	13.368
CCGAGTTA	Sample 1_Drug_Rep 1	1726277	2003125	86.18	13.298
AACTGTTA	Sample 1_Drug_Rep 1	1780094	2066669	86.13	13.342
TTGAGTAT	Sample 1_Drug_Rep 1	1855605	2162203	85.82	13.402
TTCTCAGC	Sample 1_Drug_Rep 1	1895738	2213784	85.63	13.433
CCTCCAAT	Sample 1_Drug_Rep 1	1778196	2074891	85.7	13.341
TTAGACTA	Sample 1_Drug_Rep 2	1889653	2206000	85.66	13.428
GGTCACCG	Sample 1_Drug_Rep 2	1706999	2005499	85.12	13.282
CCTCTGTA	Sample 1_Drug_Rep 2	2019307	2358690	85.61	13.524
TTGACAAT	Sample 1_Drug_Rep 2	1673842	1965360	85.17	13.253
AAGACATA	Sample 1_Drug_Rep 2	2055254	2406407	85.41	13.55
AAGATGGC	Empty well	845	5273	16.03	2.568

Typically we use Hiseq 2500, which should produce around 120-180 million reads per sequencing lane.

Unless your plate has an unusually high number of empty wells this number should be 60-90%.

Each well containing samples should be 60-90%.

Wells without template should have few matching reads.

We've noted typical ranges for the above parameters. If your data don't match the expectations noted, proceed with caution.

Inside the PoolQ Link provided to you by GPP: Counts File

PoolQ Analysis: Example_Screen

Sample
↳ Example_Screen
Clone Pool
↳ CP0041
Run Date
↳ 2020-10-08 17:30

PoolQ Input Files

	Download	Filename	Description
1	.csv	Example_Screen_Conditions.csv	CSV file mapping sample barcodes to experimental conditions
2	.csv	CP0041_reference_20160112.csv	CSV file mapping construct barcodes to barcode identifiers
3	.gz	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes
4	.gz	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes

PoolQ Output Files

	Download	Filename	Description
1	.txt	counts-Example_Screen.txt	PoolQ counts matrix
2	.txt	quality-Example_Screen.txt	PoolQ quality report
3	.txt	lognorm-Example_Screen.txt	Log-normalized counts matrix
4	.txt	correlation-Example_Screen.txt	Condition scores correlation matrix
5	.txt	barcodecounts-Example_Screen.txt	Counts by condition barcode
6	.txt	runinfo-Example_Screen.txt	PoolQ runtime information
7	.txt	data-integrity-Example_Screen.txt	Janssen discovery data integrity (DDI) document [?]

Counts File

FASTQC Reports

	Link	Download	Filename	Description
1	.html	.zip	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes FASTQC analysis
2	.html	.zip	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes FASTQC analysis

→ Go to the pooled screen analysis page

Nomenclature used in the Counts File

Sample barcode	Screening sample name	Construct barcode ID	Read count
Identifies the well of the 96-well plate. Screening samples are typically spread across many wells	Identifies your screening sample by name	Identifies individual sgRNAs	Reads detected by sequencer for each construct barcode for each sample name
(A01) TTGAACCG	Sample1_Drug_Rep1	BRDN0000946264	78
		BRDN0000921343	120
		BRDN0000944374	2
		….many more	
(A02) AATCCAGC	Sample1_Drug_Rep2	BRDN0000946264	106
		BRDN0000921343	62
		BRDN0000944374	7
		….many more	
(A03) CCGAGTTA	Sample1_DMSO_Rep1	BRDN0000946264	676
		BRDN0000921343	585
		BRDN0000944374	63
		….many more	

Total reads per **sample** should be roughly at least the minimum representation of the screen, ~500 fold more than the library i.e. ~4e7 for a Brunello screen with 77,441 unique guides.

Example Counts File

Construct barcode (sgRNA)

Construct ID

Screening sample names

Read counts

	A	B	C	D	E	F
1	Construct Barcode	Construct IDs	Sample1_Drug_Rep 1	Sample1_Drug_Rep 2	Sample1_DMSO_Rep 1	Empty Well
2	AAAAAAATCCAGCAATGCAG	BRDN0000946264	340	326	282	0
3	AAAAAACCCGTAGATAGCCT	BRDN0000921343	458	434	445	0
4	AAAAAAGAAGAAAAAACCAG	BRDN0000944374	8	61	46	0
5	AAAAAAGCTCAAGAAGGAGG	BRDN0000892836	58	325	96	0
6	AAAAAAGGCTGTAAAAGCGT	BRDN0000969073	295	214	240	0
7	AAAAAAGGGCTCCAAAAAGG	BRDN0000901979	333	301	466	0
8	AAAAACAACACATCAGAGCG	BRDN0000953921	319	363	148	0
9	AAAAACTCTGGGAAATGACT	BRDN0000951302	301	472	378	0
10	AAAAAGACAACCTCGCCCTG	BRDN0000868105	660	504	397	0
11	AAAAAGAGCTGTTTGAACAA	BRDN0000920337	758	566	547	0
12	AAAAAGATCATGATTGAGCG	BRDN0000888744	433	555	376	0
13	AAAAAGCCTGGATAACGGAA	BRDN0000882112	154	115	56	0
14	AAAAAGCTGGGTTAGAAGCG	BRDN0000893311	432	456	434	0
15	AAAAAGCTTCCGCCTGATGG	BRDN0000736924	954	1066	598	0
16	AAAAATCCGGACAATGGTGG	BRDN0000947477	269	475	311	0
17	AAAAATCCTAAAATAAAATA	BRDN0000958591	575	646	358	0
18	AAAAATCCTCTGAAGCCGCA	BRDN0000959916	202	274	320	0
19	AAAAATGCCAAAAAGCAAGC	BRDN0000970883	396	168	191	0
20	AAAAATGCGCAAATTCAGCG	BRDN0000931583	470	163	115	0
21	AAAAATGCTGAATTTCCCAG	BRDN0000863920	1023	955	812	0
22	AAAAATGGTTGACCTCACCC	BRDN0000916888	273	545	296	0
23	AAAAATGTATACTAACCAGG	BRDN0000932094	396	480	180	0
24	AAAAATTGAGAGTAACACAG	BRDN0000877631	302	303	202	0
25	AAAACAATCTCACCTCTGGG	BRDN0000945838	180	153	122	0

Inside the PoolQ Link provided to you by GPP: Lognorm File

PoolQ Analysis: Example_Screen

Sample
└→ Example_Screen
Clone Pool
└→ CP0041
Run Date
└→ 2020-10-08 17:30

PoolQ Input Files

	Download	Filename	Description
1	.csv	Example_Screen_Conditions.csv	CSV file mapping sample barcodes to experimental conditions
2	.csv	CP0041_reference_20160112.csv	CSV file mapping construct barcodes to barcode identifiers
3	.gz	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes
4	.gz	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes

PoolQ Output Files

	Download	Filename	Description
1	.txt	counts-Example_Screen.txt	PoolQ counts matrix
2	.txt	quality-Example_Screen.txt	PoolQ quality report
3	.txt	lognorm-Example_Screen.txt	Log-normalized counts matrix
4	.txt	correlation-Example_Screen.txt	Condition scores correlation matrix
5	.txt	barcodecounts-Example_Screen.txt	Counts by condition barcode
6	.txt	runinfo-Example_Screen.txt	PoolQ runtime information
7	.txt	data-integrity-Example_Screen.txt	Janssen discovery data integrity (DDI) document [?]

Lognorm File

FASTQC Reports

	Link	Download	Filename	Description
1	.html	.zip	CEA2GANXX.1.1.fastq.gz	Sequencing data file containing construct barcodes FASTQC analysis
2	.html	.zip	CEA2GANXX.1.barcode_1.fastq.gz	Sequencing data file containing sample barcodes FASTQC analysis

→ Go to the pooled screen analysis page

What values are in a Lognorm file?

$$\text{Log}_2\text{-normalized-reads-per-million for an sgRNA (log2RPM of construct barcode)} = \text{Log}_2\left(\frac{\text{Reads for the construct barcode from the counts file}}{\text{Total reads for the screening sample summed from all construct barcodes in the counts file}} \times 1e6 + 1\right)$$

1 is added in case there are 0 reads for the construct barcode

If your samples span multiple PCR plates, **DO NOT use the provided Lognorm files (these are per plate). Instead, calculate Lognorms on your own using the above formula **AFTER** summing total reads for each construct barcode across plates using the Counts File

Calculating Log Fold-Changes
(your "hit" score)

Log Fold-Change (LFC) = log2RPM of your *experimental* sample − log2RPM of your *reference* sample

Discussion point: What is your *reference* sample? Is it an early time point? pDNA as a stand-in for an early time point? A late time point of a sample not treated with drug? These are all examples of possible reference samples that could be used here.

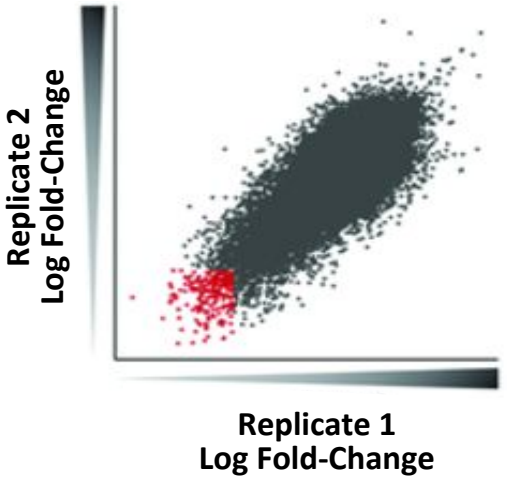
- Note that a simple Log Fold-Change calculation is only a first-pass of analyzing the data, and more sophisticated methods maybe called for to determine a 'final' hit-list.

Graphs you should look at to determine the quality of your data

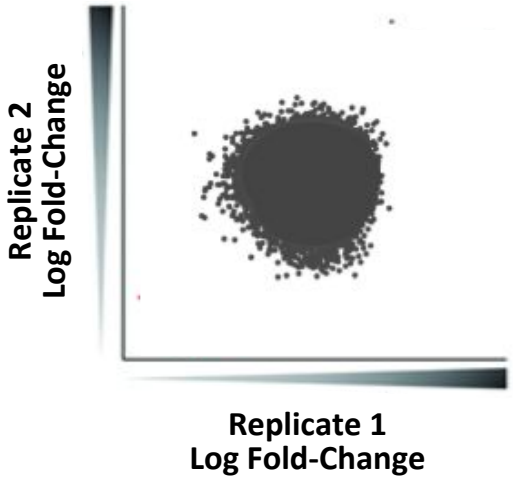
Understanding replicate correlation graphs of Log Fold-Changes

Focus on Negative-Selection Outcomes

Beautiful screen for negative (as well as positive) selection



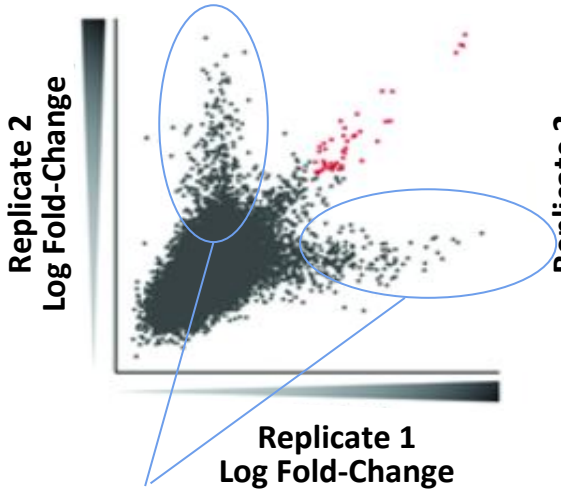
Screen with issues



General lack of correlation or outliers indicates lack of signal

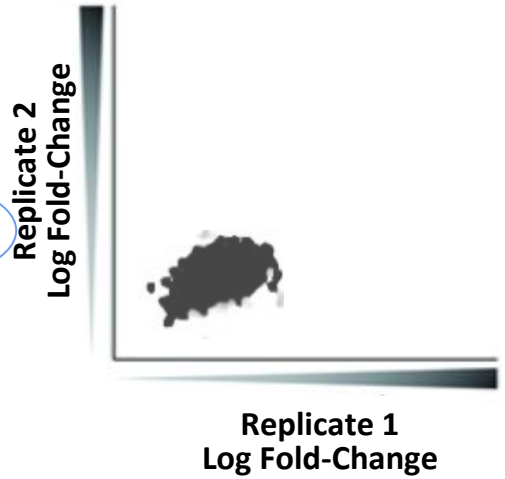
Focus on Positive-Selection Outcomes

Nice drug-resistance screen



Presence of these points indicate a modest lack of replicate reproducibility, but reproducible hits are still apparent on the diagonal

Drug-resistance screen with issues



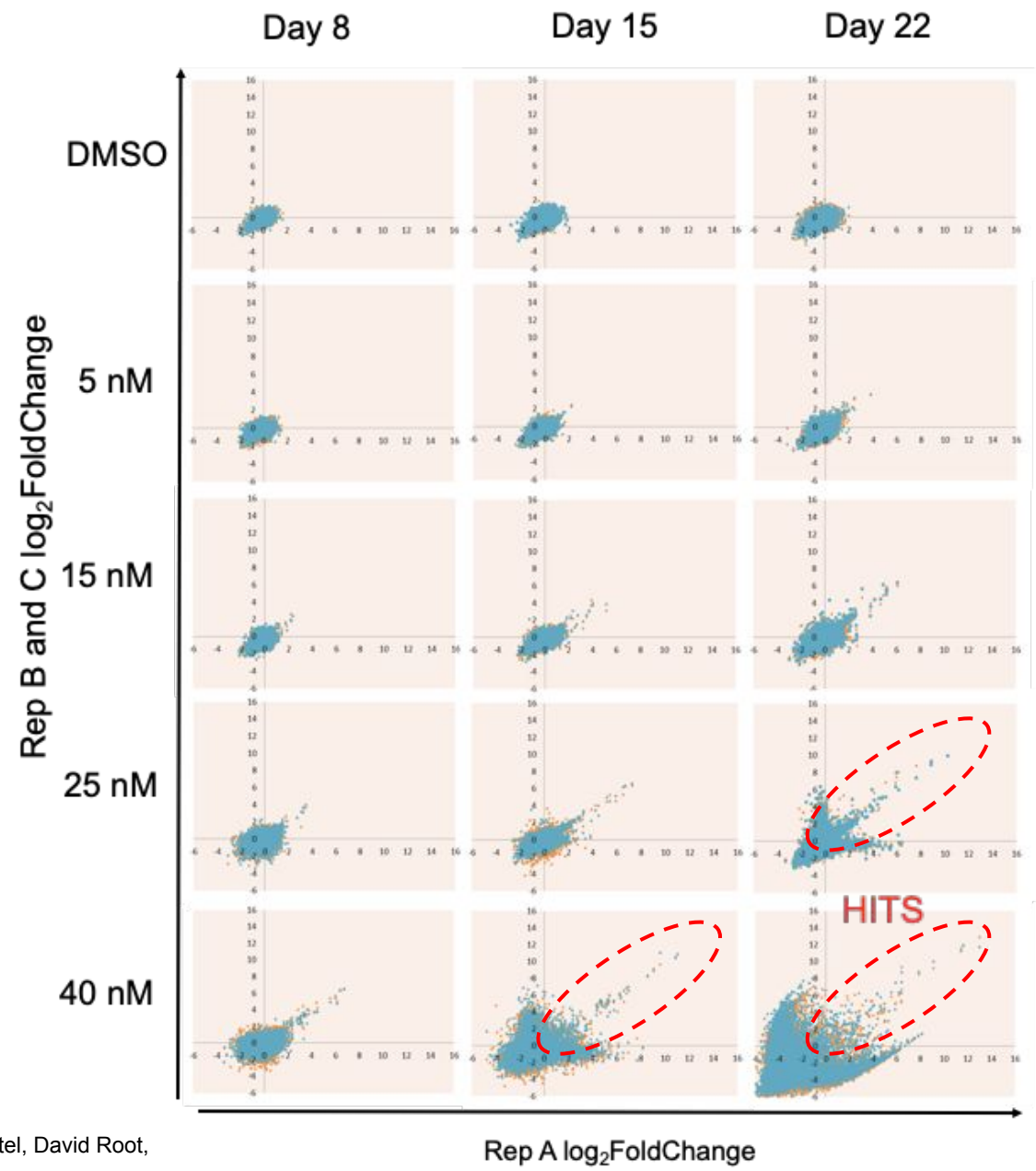
Lack of outliers indicates not enough drug was used, or not enough time passed (see following slides)

Check all replicates against each other;
if one replicate consistently looks bad consider dropping it from analysis.

Pictures modified from:

Piccioni F, Younger ST, Root DE, *Curr Protoc Mol Biol*, 2018 Jan 16;121:32.1.1-32.1.21 and Doench JG, Hanna RE, *Nat Biotechnol*, 2020 Jul;38(7):813-823

The effects of drug concentration and length of screen on signal for drug resistance, focusing on positive-selection outcomes



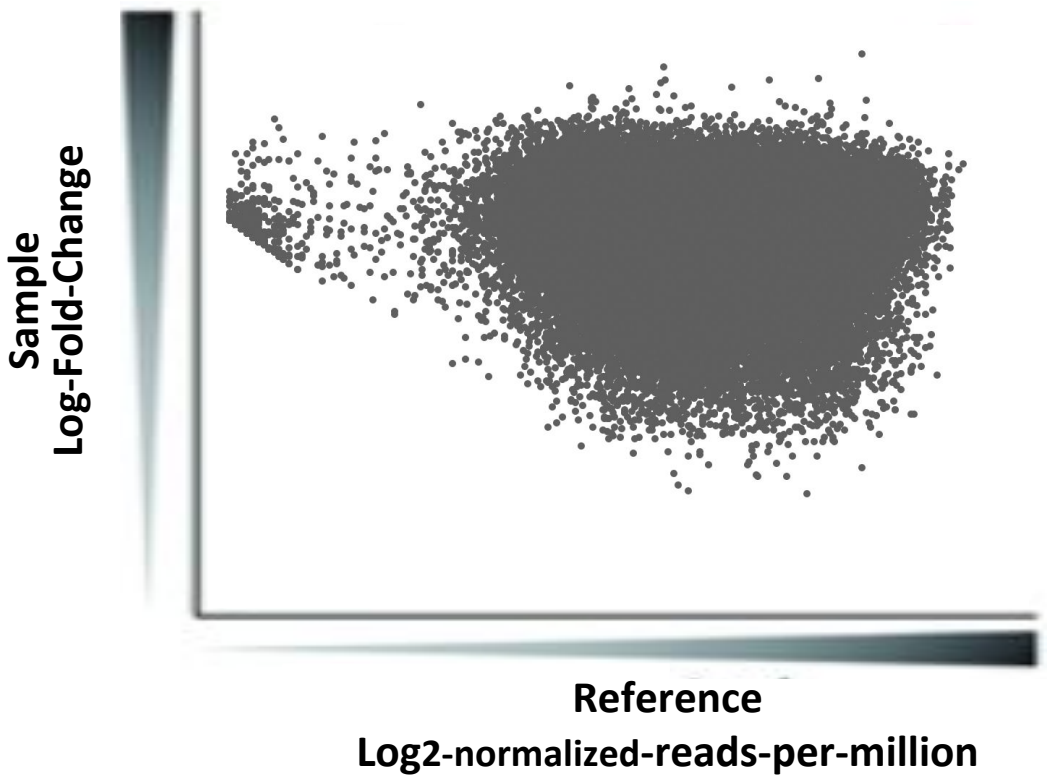
Drug concentration and length of screen are quite important for the strength of your hits. You want to have used enough drug and for the screen to have progressed for enough time.

If you collected pellets at multiple timepoints, and you sequenced too early (not enough signal) or too late (too much noise), you can sequence the other timepoints and see if you did better.

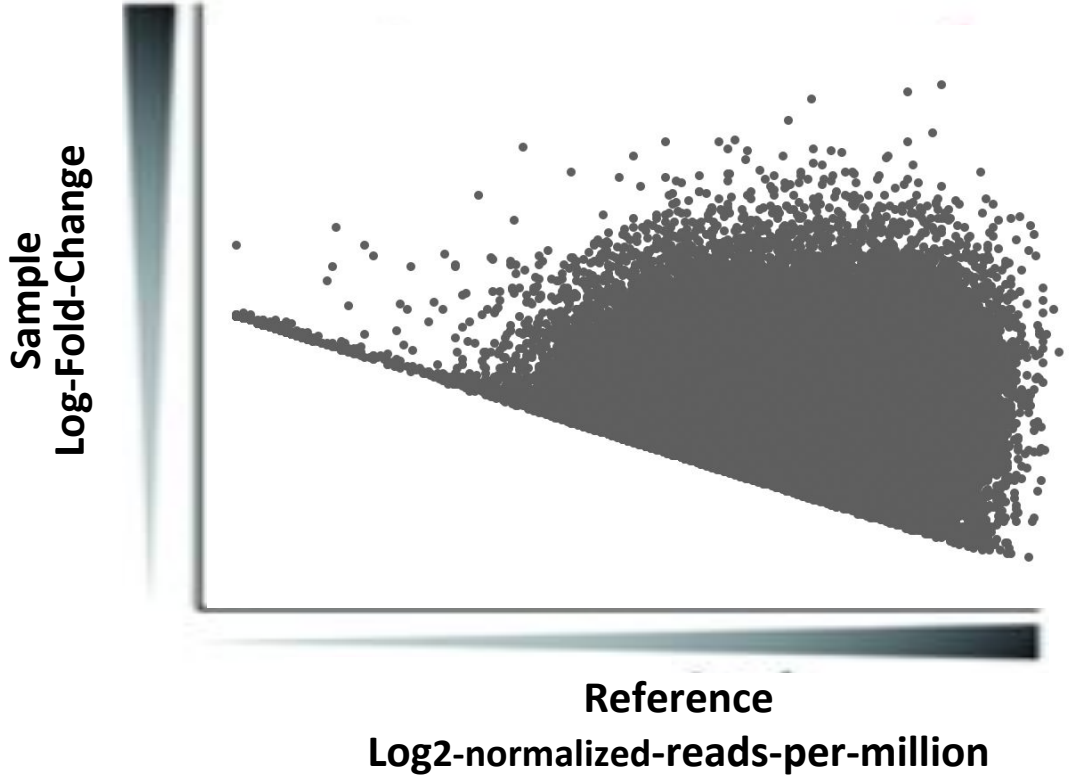
Meghan Wyatt, Amy Goodale, Yenarae Lee, Ting Wu, Sasha Pantel, David Root, Cory Johannessen, Federica Piccioni

Analyze sample Log Fold-Changes relative to your reference representation

It's pretty good



It's not so great



Buildup along a straight line in this way indicates that some guides don't have enough representation to measure negative-selection.
This is evidence of a bottleneck in your screen.
Some positive-selection hits may still be detected.

Calling your hits: basic guidelines

How to call hits is a quite important and complicated question, which we can't fully answer. Here we will provide some guidance and rules of thumb, specifically for CRISPR pooled-screening of commonly used libraries.

From Guides to Genes for CRISPR screens

We recommend generating a volcano plot as a first-pass:

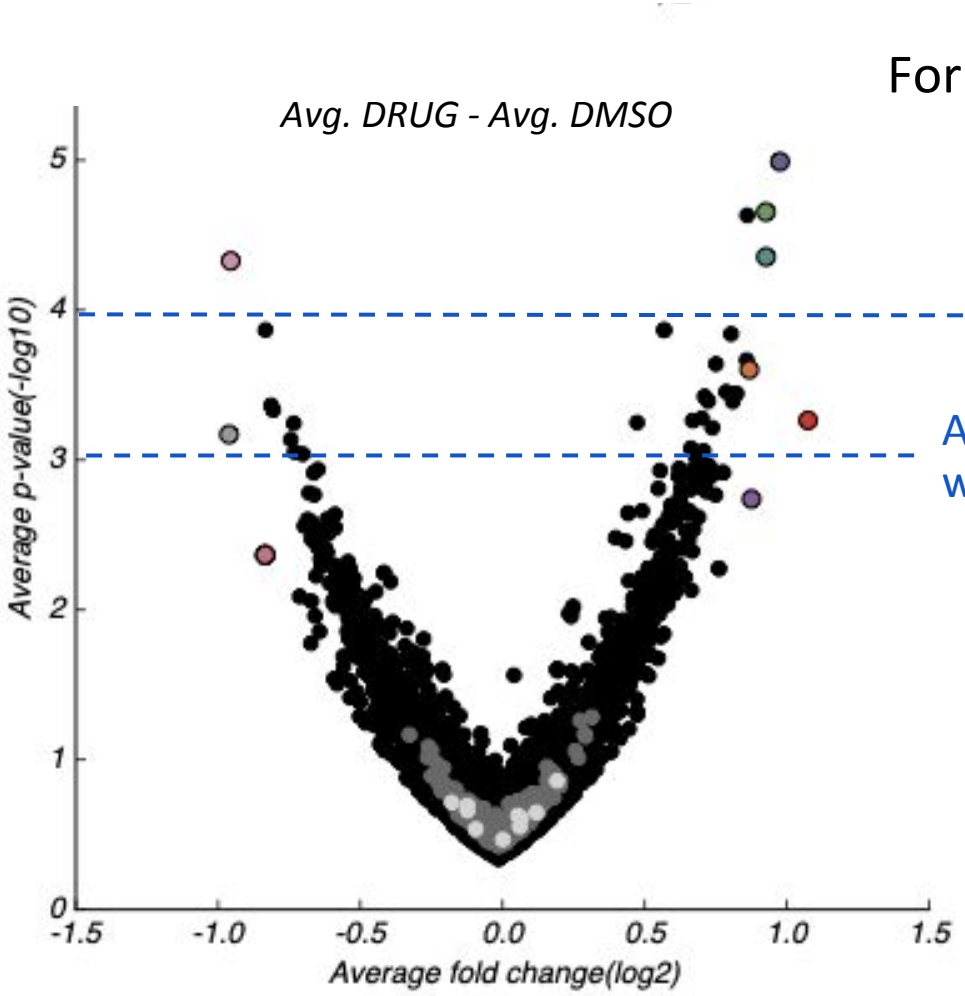
https://portals.broadinstitute.org/gpp/public/analysis-tools/crispr-gene-scoring

This will generate volcano plots and gene-level-scores using the CHIP file and Log Fold-Changes values, provided in **.txt file** form. There are many other ways to analyze screens, and multiple methods can be compared to ensure that results are robust to the exact analytic approach.

See: Doench JG, Hanna RE, *Nat Biotechnol*, 2020 Jul;38(7):813-823 for discussion of other available tools

Volcano plots of *p*-values vs. Log-Fold-Change by gene:

What do we consider great, ok, and less-than-great hit values for some standard libraries?



For a standard Brunello or Avana **4 guides per gene** CRISPR knock-out screen:

Anything above 4 is likely to validate

Anything above 3 is potentially worth follow-up

Anything below 3 should be viewed with skepticism

The hypergeometric tool on our website calculates *p-values* based on the consistency of ranks of guides targeting the same gene.

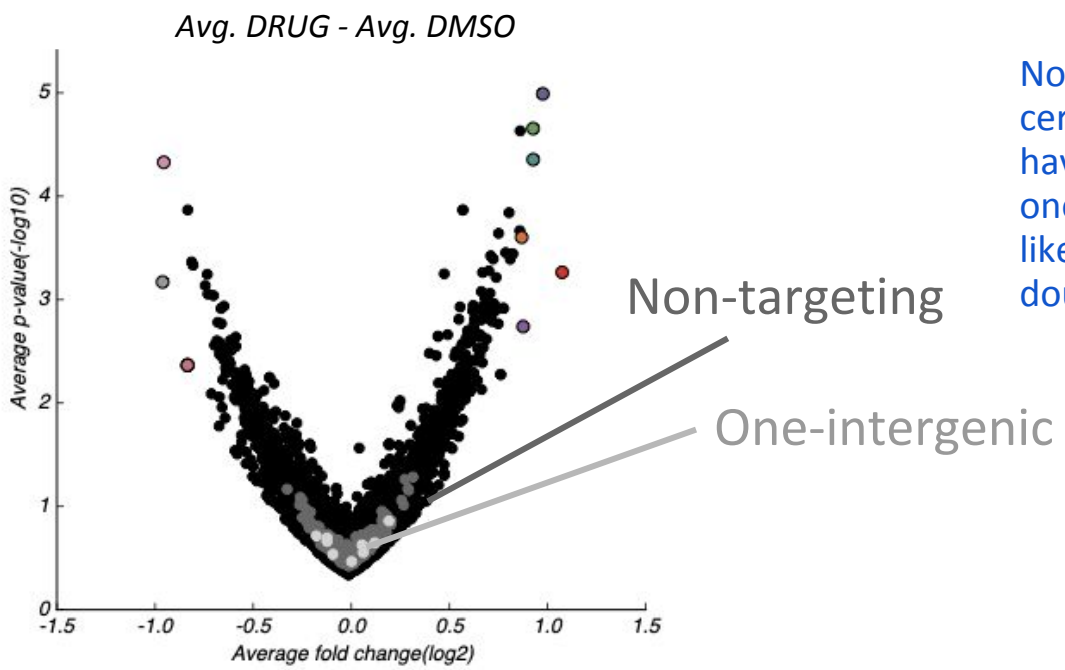
For screens without multiple guides (like ORF screens), the reported value is not applicable and Log Fold-Change alone might be the better way of determining a hit cutoff.

Determining a hit cutoff based on Log Fold-Change values alone is best done with proper use of control constructs, to establish an empirical null.

What are the Log Fold-Changes of your negative and positive controls?

For a standard CRISPR knock-out screen:

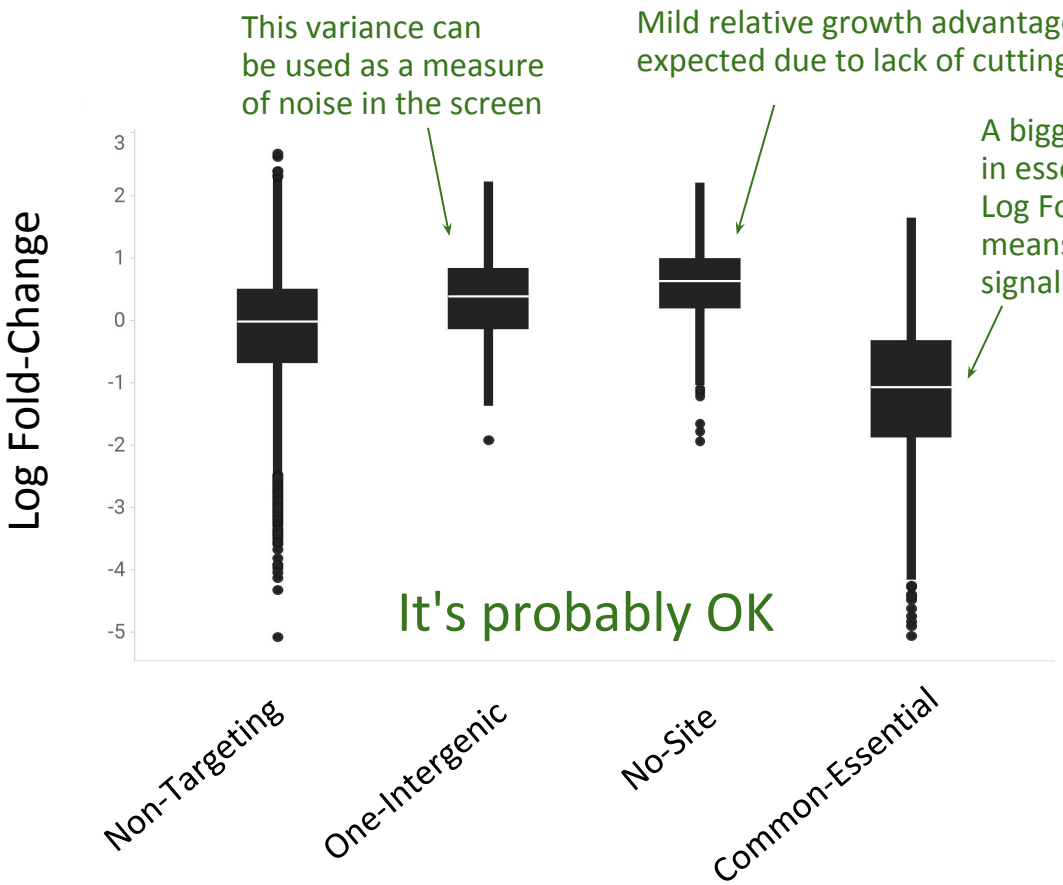
Non-targeting (NO_SITE) and One-intergenic-site (ONE_SITE) control guides can give an indication of screen noise. These should be much lower Log Fold-Change and *p*-value than your hits.

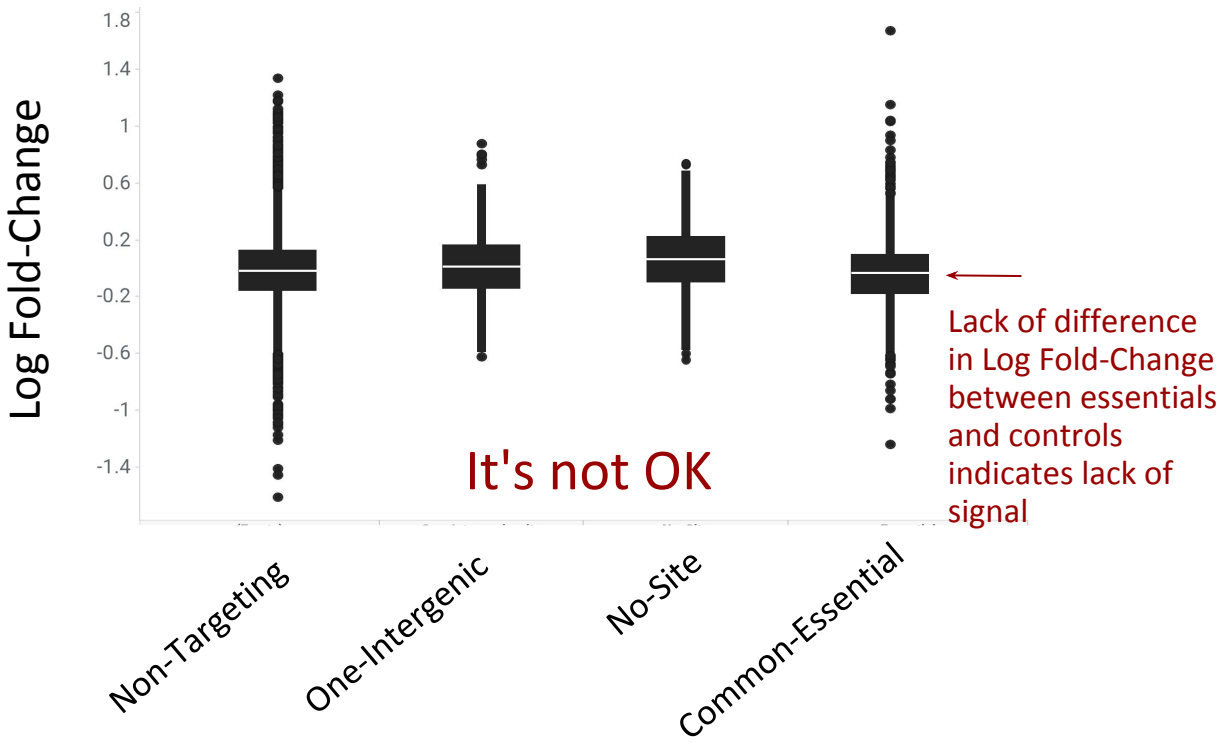


Note: it is not uncommon that for certain cell lines non-targeting controls have a relative growth advantage over one-intergenic site controls due, most likely, to increased sensitivity to double-strand breaks.

What are the Log Fold-Changes of your negative and positive controls?

For a standard CRISPR knock-out, negative-selection screen



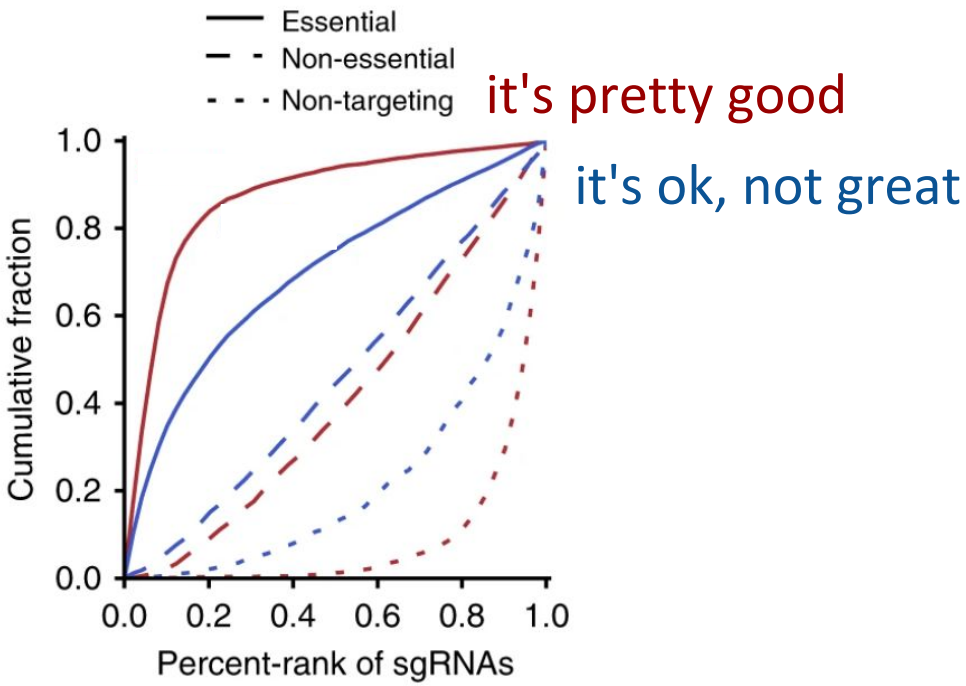


The gold-standard set of common essential gene is from: Hart T et al Cell 2015;163:1515–1526 and Hart T et al Mol. Syst. Biol. 2014;10:733 (this list is also available at https://github.com/mhegde under auc-calculation)

What are the Log Fold-Changes of your negative and positive controls?

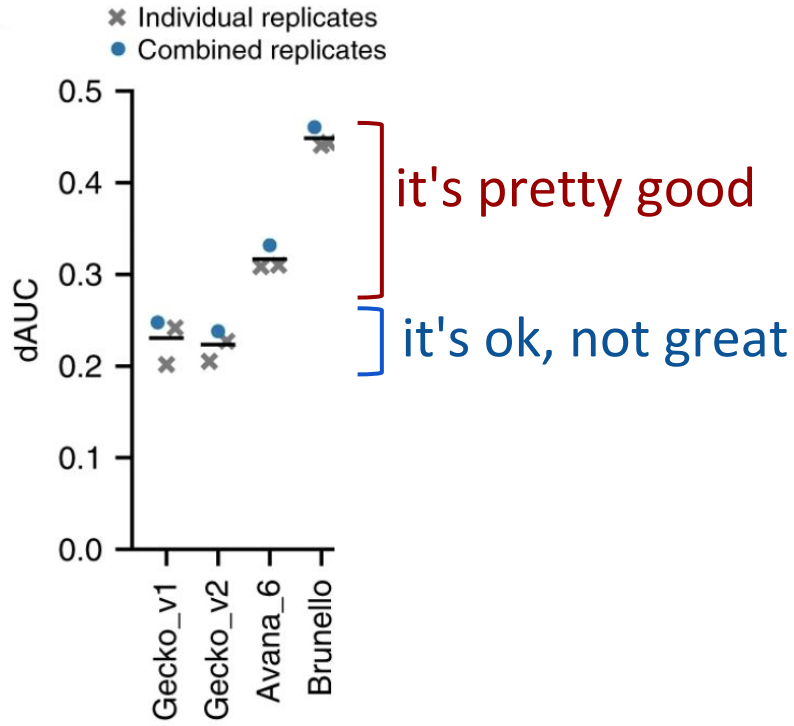
For a standard CRISPR knock-out, negative-selection screen

Interpretation of common essentials and common non-essentials allows your screen to be compared quantitatively to other screens



it's pretty good

it's ok, not great



it's pretty good

it's ok, not great

The dAUC can be calculated as described in Sanson KR et al. *Nat Commun*, 2018 Dec 21;9(1):5416

For reference and more in depth discussion see literature published from the GPP regarding general pooled-screening design and analysis:

Piccioni F, Younger ST, Root DE, *Curr Protoc Mol Biol*, 2018 Jan 16;121:32.1.1-32.1.21

Doench JG, *Nat Rev Genet*, 2018 Feb;19(2):67-80

Doench JG, Hanna RE, *Nat Biotechnol*, 2020 Jul;38(7):813-823

For further aid with analysis for your specific project please contact your screening scientist.