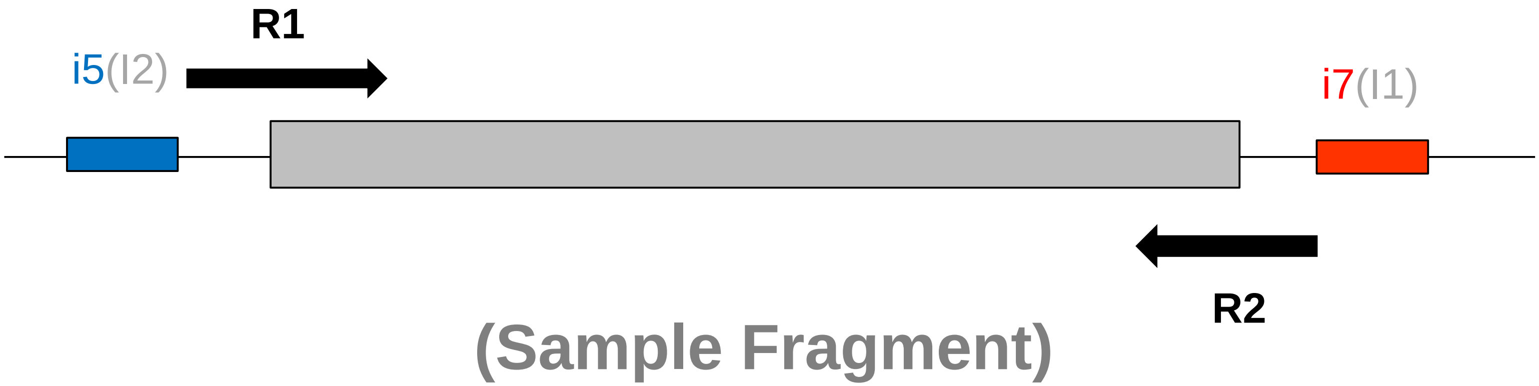
# Effect of Calling Single-Barcode Samples as Dual-Barcode Samples

Charles Warden

Integrative Genomics Core

**(idea from Jinhui Wang and Xiwei Wu)**

# Illumina Barcode Design



- Single-Barcodes only use i7
- Dual-Barcodes also use i5 (regardless of whether single-end **R1** or paired-end **R1+R2**)
  - You can call single-barcode samples as dual-barcode samples, based upon the adapter (for most samples, TCTTTCCC can be used for the 2nd barcode, i5)
- Diagram similar to what Jinhui drew for me

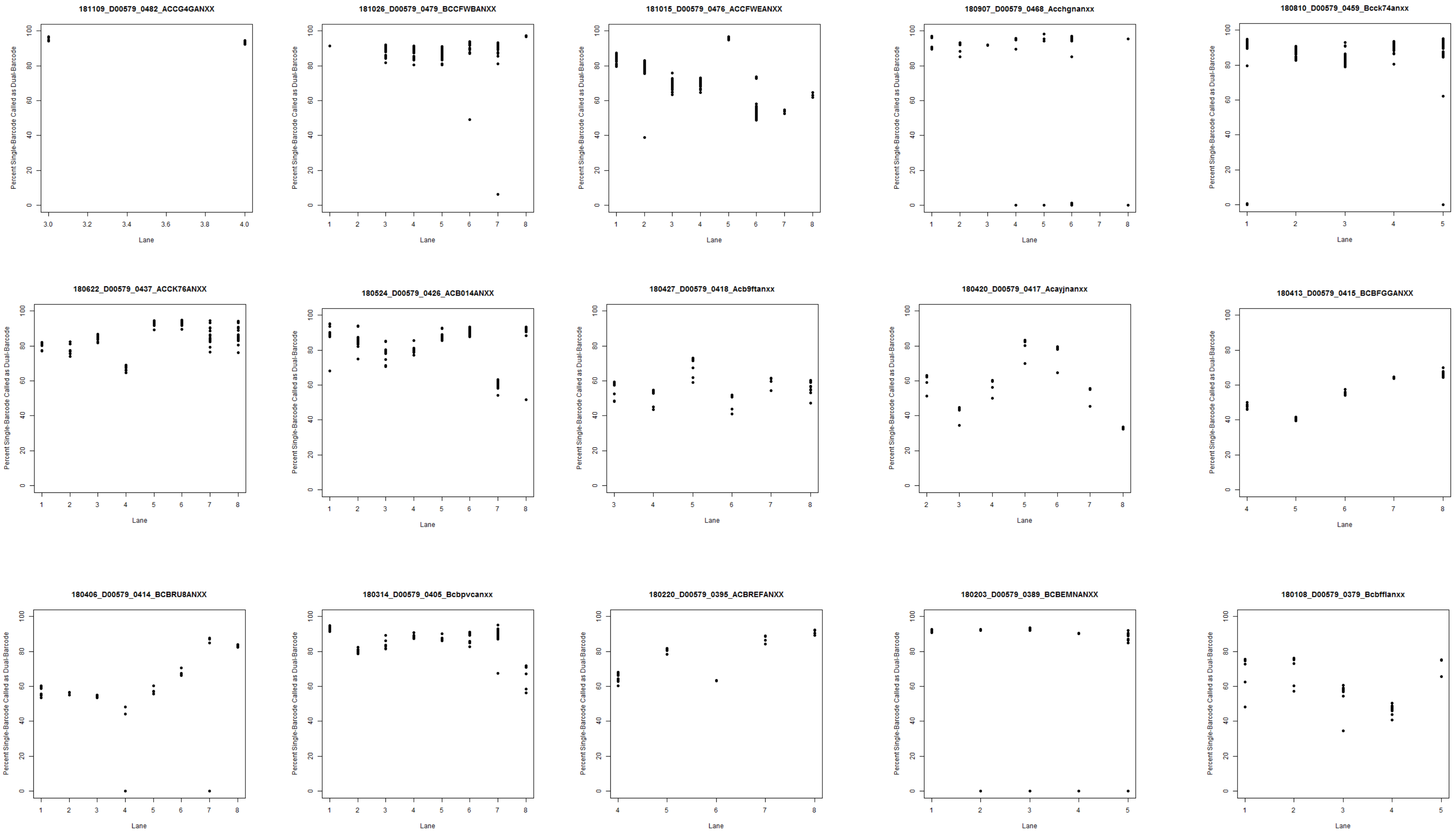
# Experimental Design

- Look at runs with a mix of single-barcode samples and dual-barcode samples, for **paired-end (PE)** libraries
  - This makes up 16 / 99 runs between 1/8/2018 and 11/9/2018
    - Also included 2 all single-barcode runs (that were a mix of regular and 10X samples)
  - So, this is a subset of runs (Rapid Run and Regular Run) from our 2 HiSeq 2500 sequencers
- Assuming 2nd barcode should always be TCTTTCCC for single-barcode samples, test calling single-barcode as dual-barcode samples
  - **Rationale**: we happened to do this for a run, and noticed a decrease in the reads assigned with dual-barcode versus single-barcode.
  - So, we could test base calling "**single-barcode as dual-barcode**" for 2,129 PE "samples" (from 18 runs)
    - Some samples are split between lanes (technical replicates with separate lane effects)
      - 1,793 *unique* sample IDs for SB samples (which can be tested for calling as dual-barcode samples)
      - 1,400 *unique* sample IDs for DB samples
- Just in case the trend was different for **single-end (SE)** runs, I also went back and analyzed 5 other runs during that same time interval.

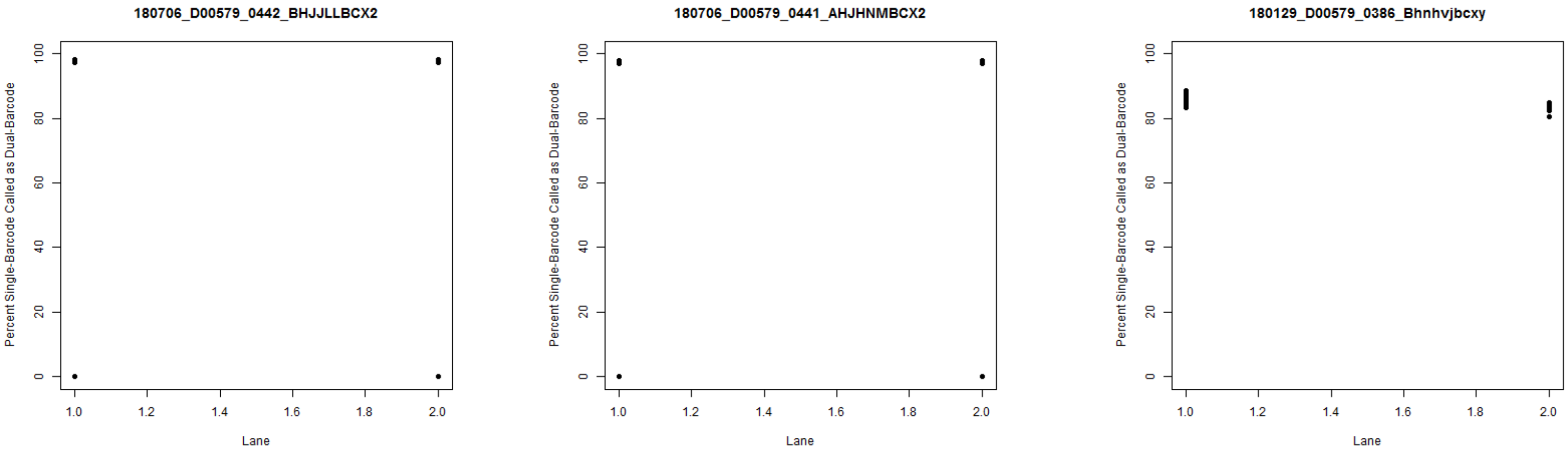
# Question #1a

**Do Paired-End (PE) single-barcode sample typically have ~15% loss as dual-barcode samples?**

# Runs Usually Have (Some) Noticeable Lane Effects (PE Runs)

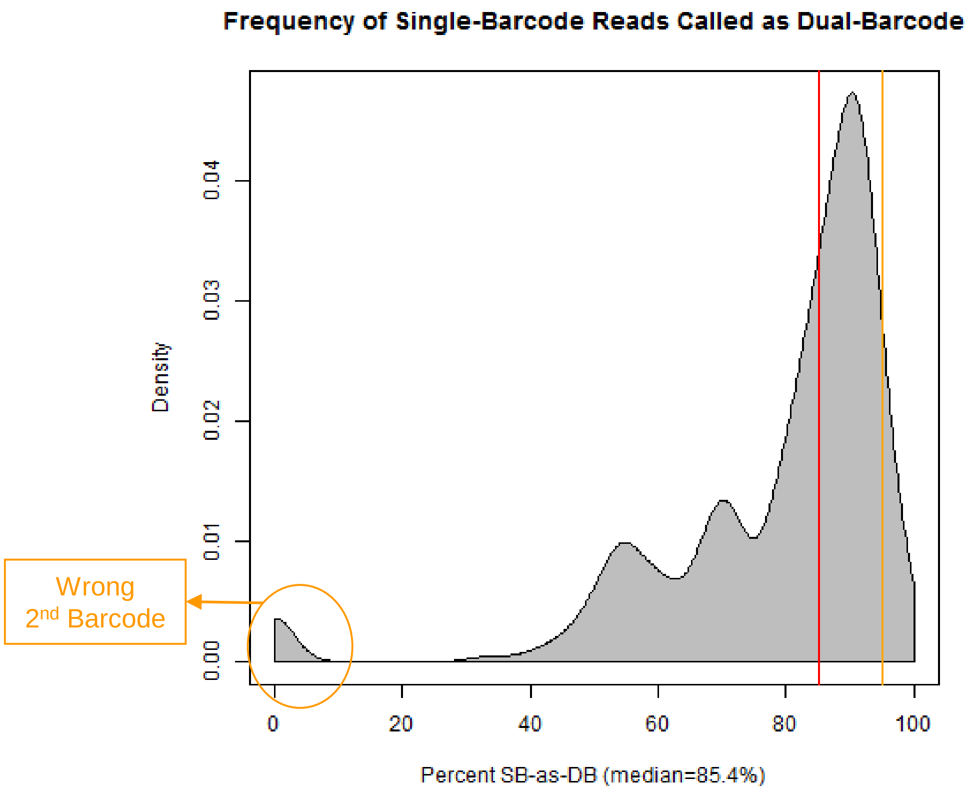


# Having Fewer Total Samples Helps with Lane Effects? (Only Rapid PE Runs Shown Below)

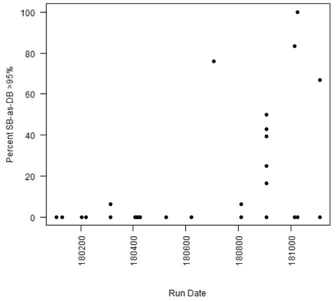


- Small number of runs, but there aren’t any intermediate SingleBarcode-as-DualBarcode percentages (not between 20% and 80%)
- As indicated in “Additional Information,” all <2% samples can be explained by incorrect 2nd barcode (for single-barcode samples prepared with a dual-barcode kit, but didn’t require 2nd barcode when originally returned).

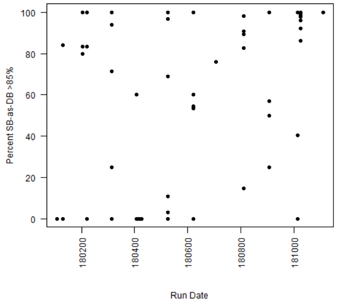
# Frequency of Read Loss as Dual-Barcode (PE Runs)



**>95% Kept (per lane/run)**



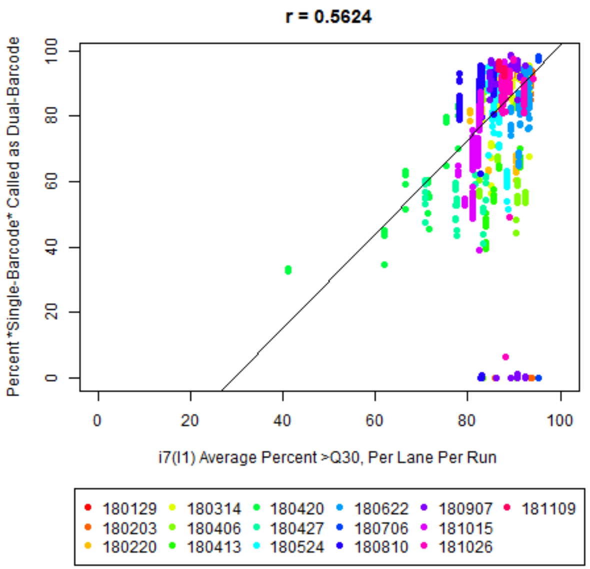
**>85% Kept (per lane/run)**



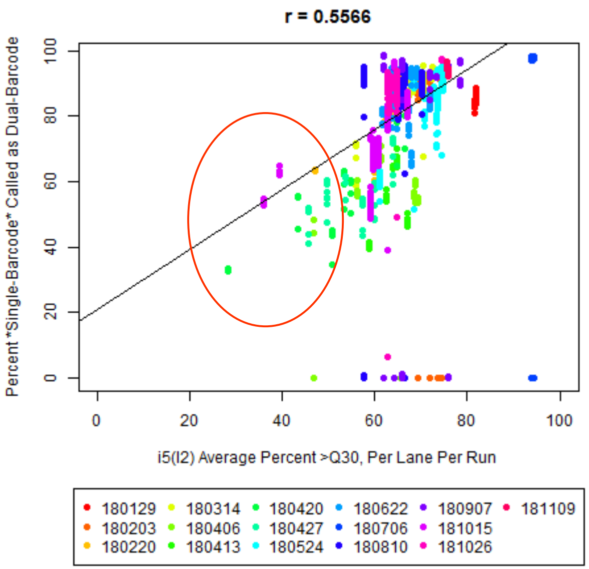
- Mapping of correct barcode for “Wrong 2nd Barcode” samples is explained on Supplemental Slide #22
- Grey shade is for distribution across all samples
- I always ran bcl2fastq (v2.18) base calling with --barcode-mismatches 0, and I believe that is also what was done for all runs presented here (I always performed dual-barcode base calling for these single-end samples, and I sometimes performed the original single-barcode .fastq generation)
- **Almost all runs would fail to have >95% kept, but much larger fraction of runs can have most samples with >85% of reads kept**

# Effect of Quality Scores on Percent SB-as-DB (PE Runs)

## i7(I1)

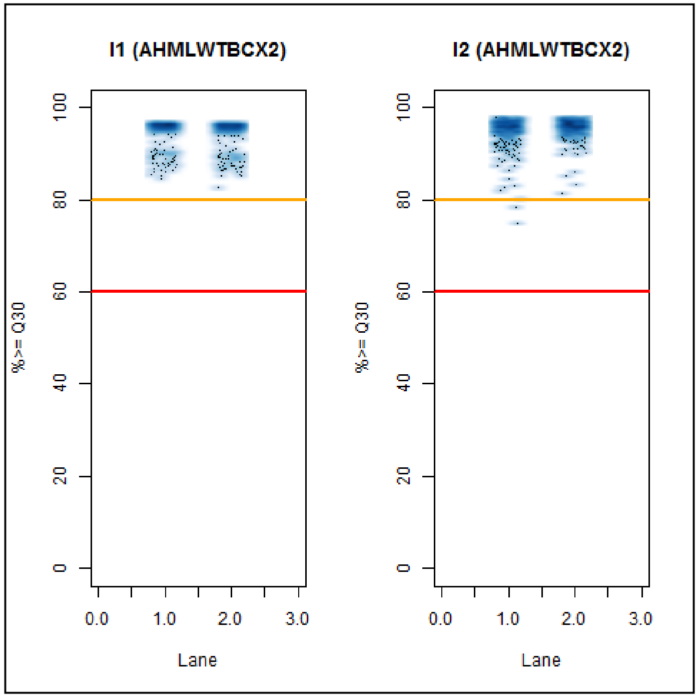


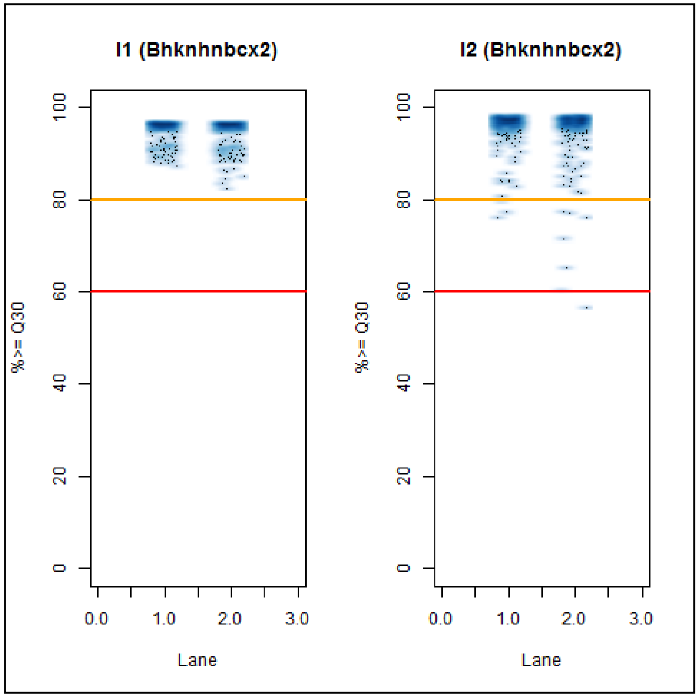
## i5(I2)



- If outliers are removed, the drop in I2 to Average Percent >Q30 identifies many of the runs with SB-as-DB percentages between 20% and 80%
  - Correlation and linear regression trendline only calculated with samples whose Percent SB-as-DB is >10%
  - **Extra testing performed (mostly not shown), but these seem to be best fit among remaining samples (out of what I have checked)**
    - I think the 2nd best fit was for the Percent PF (slide #30, left-side), but that can decrease in some samples with a higher Percent SB-as-DB
  - **11.4% of samples would be flagged with Percent SB-as-DB values between 20% and 60%** (roughly the range that includes the lower I2 Percent Q30 for a subset of samples)
- As expected, percent >Q30 is higher for i7(I1) than i5(I2)

# Example of % >Q30 for More Recent 2019 Run (HiSeq Rapid Run, PE Libraries)

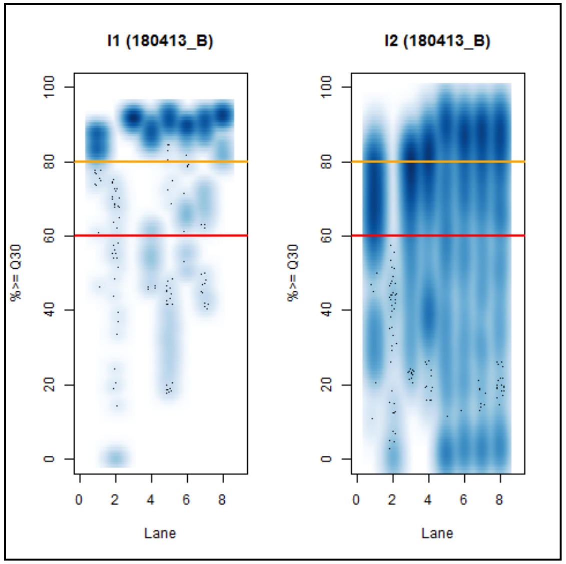


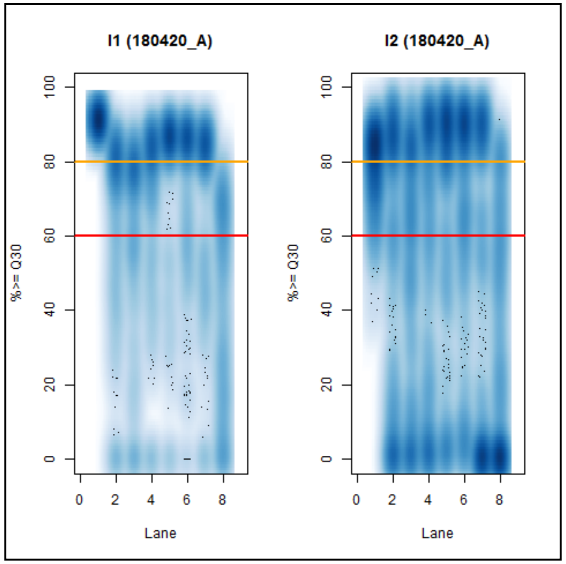


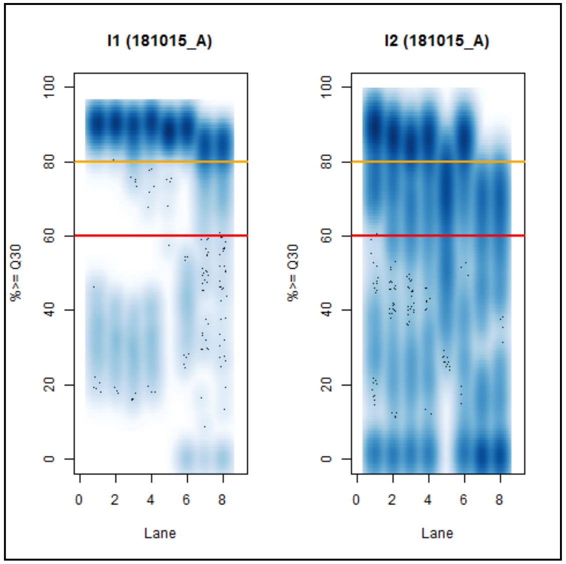
- I1 is i7 index, I2 is i5 index
- With this paired-end **Rapid Run** (with a mix of single-barcode and dual-barcode samples), notice that all tiles are above the range where we noticed read loss with dual-barcodes on the previous slide
  - Previous slide is the average value from this sort of plot (per-lane), which is ~95% for this run
  - Samples combined between lanes and flow cells (so, right and left plots are technical replicates, as well as lanes within each plot)
  - So, let’s consider this an example of a “good” run

# Example of 3 Runs With Low Percent >Q30 for I2

## (Selected 2018 PE Runs, Low % >Q30 in at least some lanes)





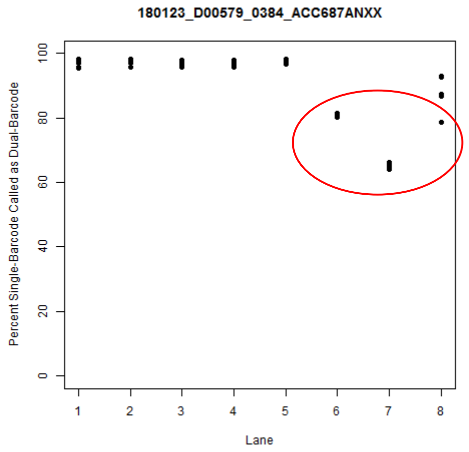


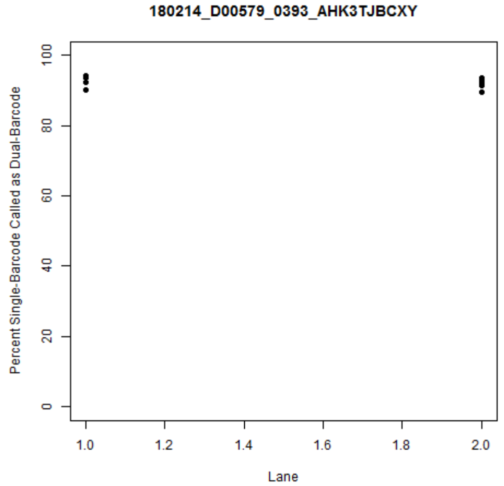
- I would recommend creating these figures after the initial base calling, before returning reads to users and summarizing reads actually created in their folder
- Notice this is substantially different than what I would consider to be a "good" run on the previous slide
  - The density for a substantial fraction of samples has to drop down considerably for the average I2 Q30 to decrease (like in **slide #8**)
  - I think users should at least be warned the next time a run like this is encountered

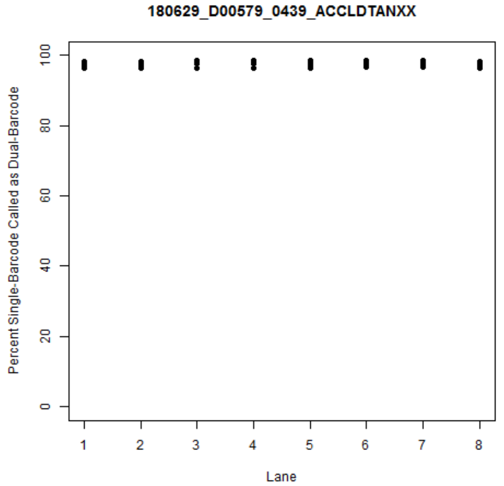
# Question #1b

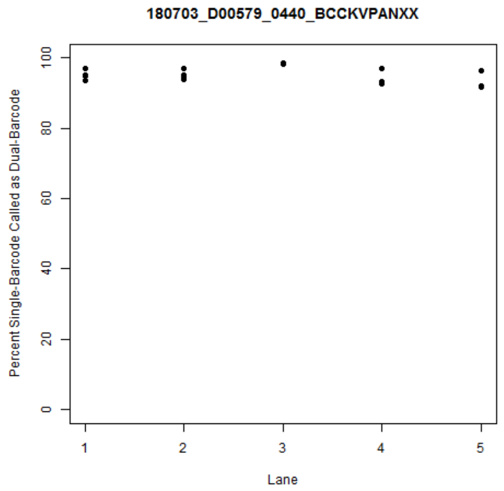
**What happens to the SingleBarcode-as-DualBarcode rate when Single-End (SE) samples are used instead of Paired-End (PE) samples?**

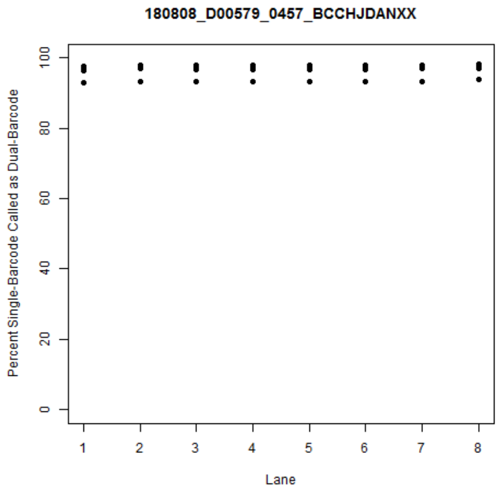
# Lane Batch Effects Improve for Single-End Libraries







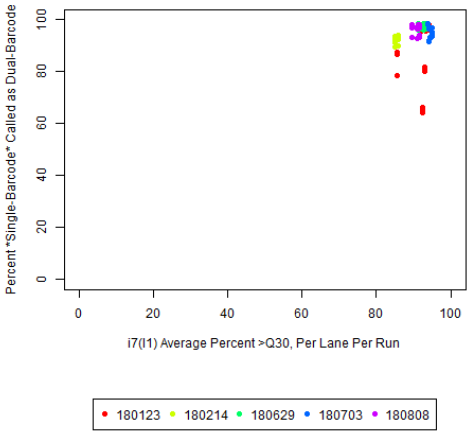




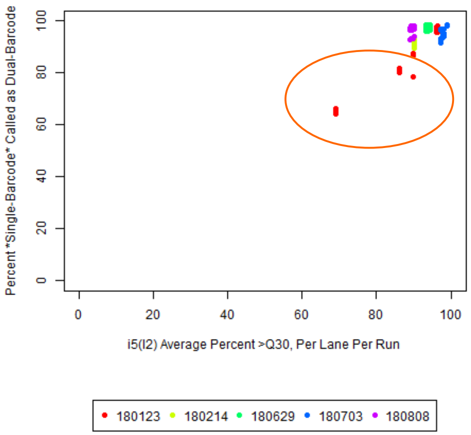
- I initially skipped one run (180123) due to large number of sample description files in folder (and concern about sample mapping)
- However, if you consider the 4 runs initially tested, they all have SingleBarcode-as-DualBarcode percentages greater than 80% for single-barcode libraries
  - While number of runs is considerably less, I would guess the dual-barcoding to be a bigger issue for paired-end (PE) libraries than the single-end (SE) libraries (at least in terms of the effect of I2 quality)

# Effect of Quality Scores on Percent SB-as-DB (SE Runs)

## i7(I1)



## i5(I2)



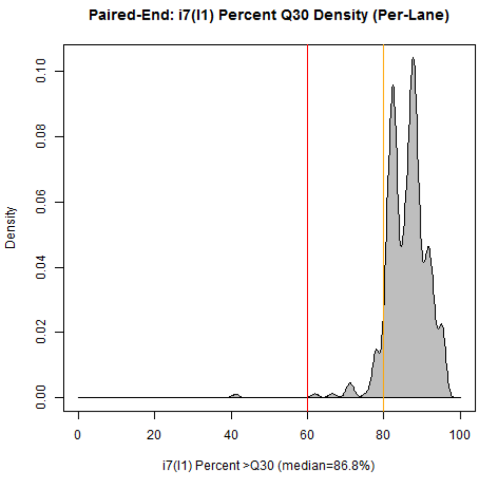
- Interestingly, the 1 out of 5 runs that is more of an outlier is the one that I waited to re-process (because the bcl folder had several CSV sample sheets, indicating considerable testing with alternative sample mapping / base calling strategies)
- Since percentage relatively high with single-end libraries, set y-axis to be scaled from 0 to 100
  - Also, remove trendline because absolute percentage matters more than correlation
  - Nevertheless, **for the remaining 4 runs**, the point is that the Q30 is relatively high for both indices (for single-end library), and the SingleBarcode-as-DualBarcode percent was >80% for all 4 of these runs

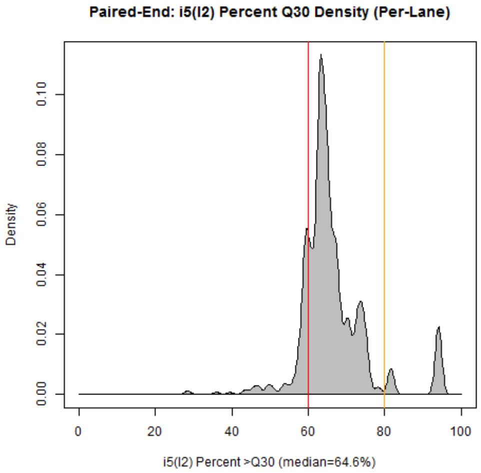
# Distribution of Percent Q30 Values

## (From Samples with Mixed Single-Barcode and Dual-Barcode Runs)

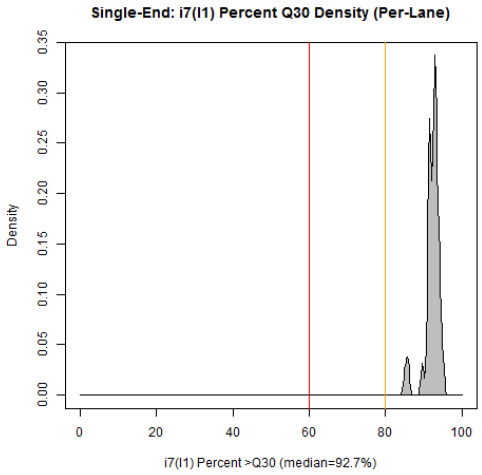
i7(I1) i5(I2)

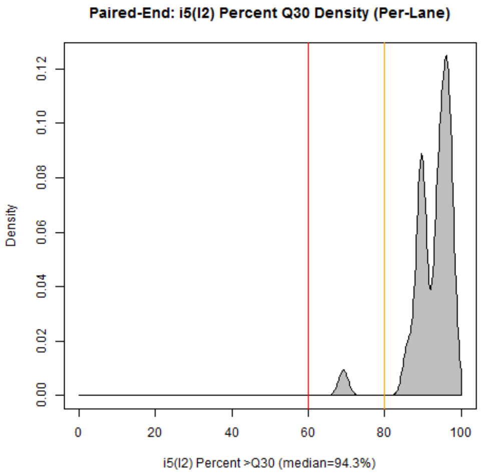
**Paired-End (PE) Libraries**


Paired-End: i7(I1) Percent Q30 Density (Per-Lane)


Paired-End: i5(I2) Percent Q30 Density (Per-Lane)

**Single-End (SE) Libraries**


Single-End: i7(I1) Percent Q30 Density (Per-Lane)


Paired-End: i5(I2) Percent Q30 Density (Per-Lane)

- As explained in the main results, single-end percent Q30 distribution was better (although, *perhaps outlier threshold should vary*, or at least there are different severity for samples below orange threshold *versus* red threshold)
- However, whether 50% or 60% should be used as the threshold for a quality flag (for PE libraries) is a little unclear:
  - 60% seems to be a better match for SingleBarcode-as-DualBarcode outliers on **slide #8**
  - However, 50-55% seems like a more appropriate threshold for the bulk of the percent Q30 values (for paired-end I2), on this slide

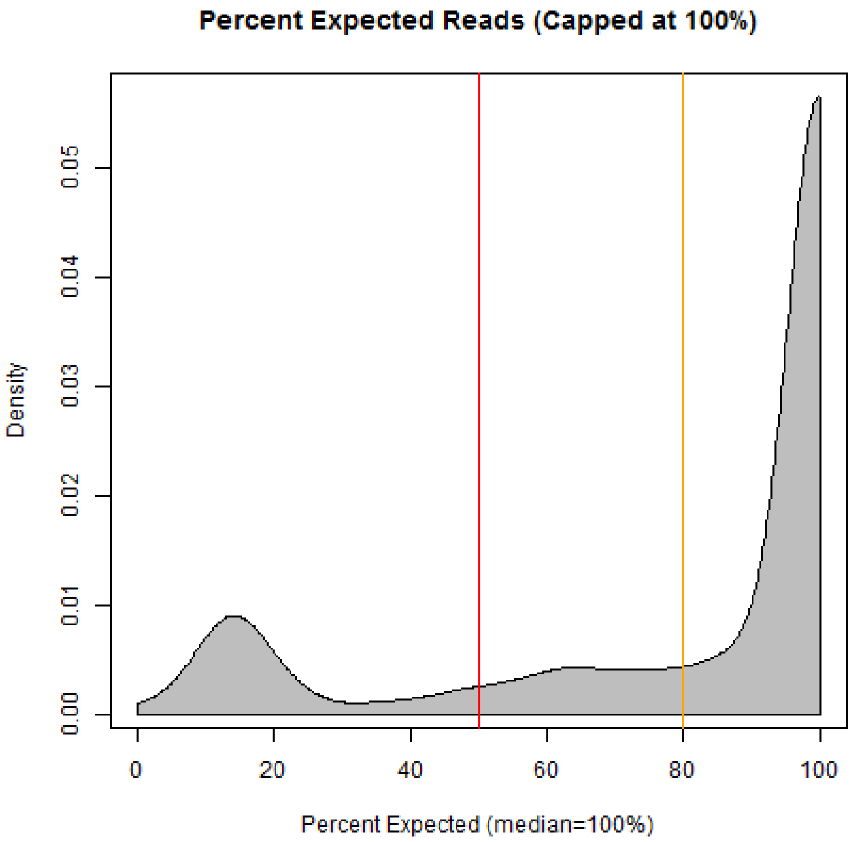
# Question #2

**Does the SingleBarcode-as-DualBarcode Percentage have anything to do with variation from the expected number of reads?**

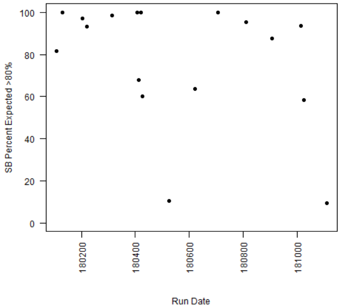
Primarily focus on samples with less than expected reads (with percent expected capped at 100%) in the following slides.

# Frequency of Meeting Expected Read Counts

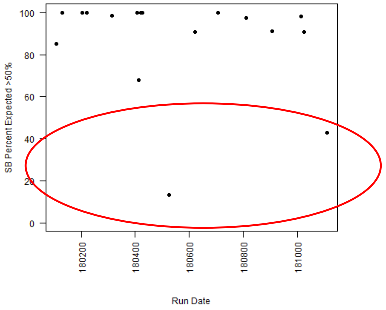
(**Single-Barcode PE Run**, Among Runs with Mix of Single-Barcode and Dual-Barcode Samples)



**>80% Expected (per lane/run)**



**>50% Expected (per lane/run)**

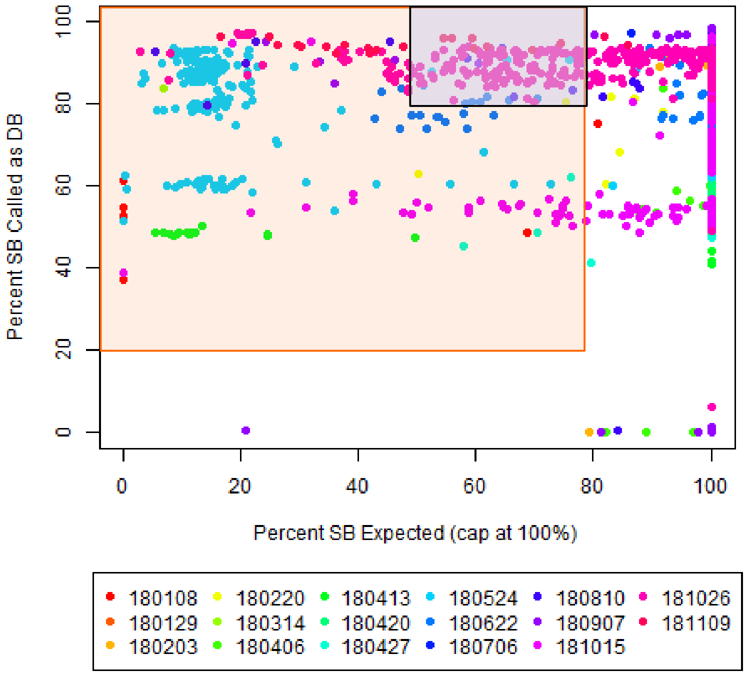


- Notice thresholds are different than single-barcode as dual-barcode (for example, a 50% read loss would typically require combining reads between runs)
  - **However, most runs have >50% of the expected reads (see right), and the median sample exceeds the number of expected reads (I think we planned for extra reads)**
- **However, this is in the ballpark of the total 15% read loss for SingleBarcode-as-DualBarcode.**

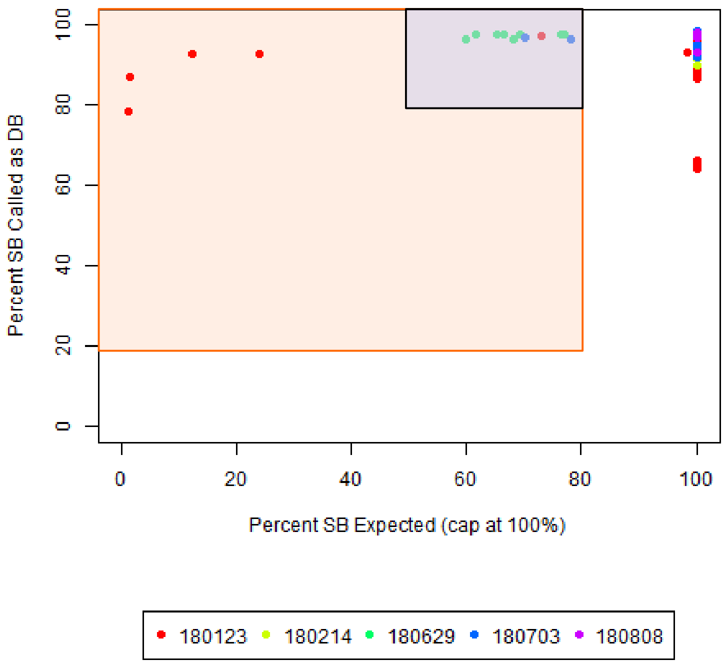
# Effect of Percent SB-as-DB on the Percent SB Expected

## (All Originally Single-Barcode Samples)

## PE Runs



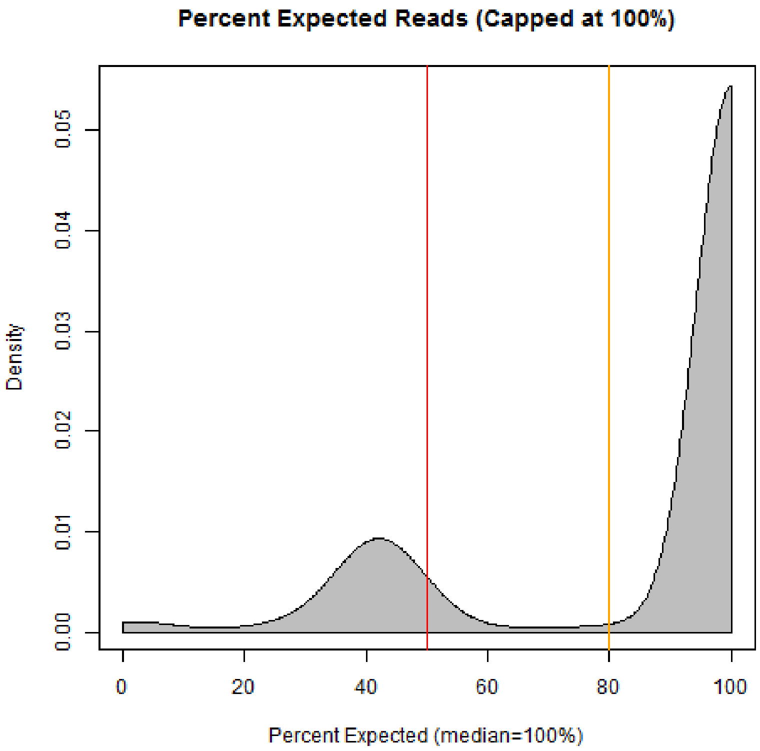
## SE Runs



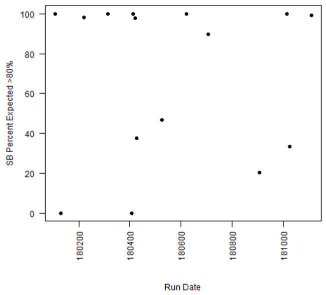
- Sample size is limited, but we no longer have lowest SB Expected or intermediate SB-as-DB with Single-End (SE) Runs (versus Paired-End Runs)
  - **So, values are not highly correlated, but there are sectors that seem to be more likely filled by Paired-End (PE) runs**
- **Also, SB-as-DB is usually a tile/lane measure, and SB Expected is usually a sample measure** (with samples commonly combined between lanes). So, *unlike most plots*, I summed the SB and DB counts to calculate an overall percent SB-as-DB value (per-sample).
- We would like to avoid having to re-process samples and combine reads between runs (or throw out runs)
  - While having more mixed SE runs is important, I believe these results are sufficient to start a discussion about having flags for either :
    - a) majority samples have <50% expected reads
    - b) low I2 Q30
  - These flags may be useful, *even if they capture different issues overall*

# Frequency of Meeting Expected Read Counts

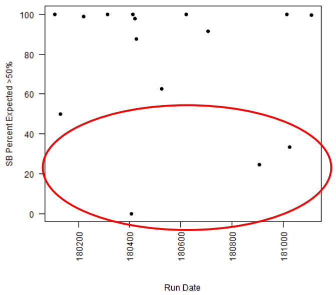
(**Dual-Barcode PE Run**, Among Runs with Mix of Single-Barcode and Dual-Barcode Samples)



**>80% Expected (per lane/run)**



**>50% Expected (per lane/run)**

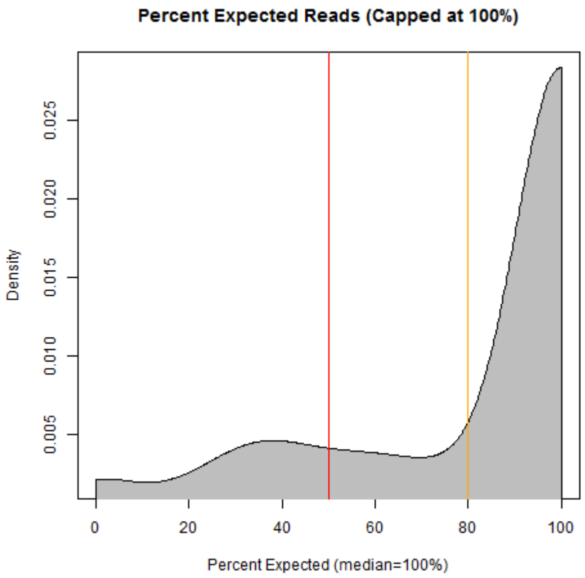


- Notice thresholds are different than single-barcode as dual-barcode (for example, a 50% read loss would typically require combining reads between runs)
  - **However, most runs have >50% of the expected reads (see right), and the median sample exceeds the number of expected reads (I think we planned for extra reads)**
- Samples with <50% samples with >50% expected reads make up 3 out of 16 runs (so, ~15-20%)
  - **Again, this is roughly in the ballpark of the total 15% read loss for SingleBarcode-as-DualBarcode.**

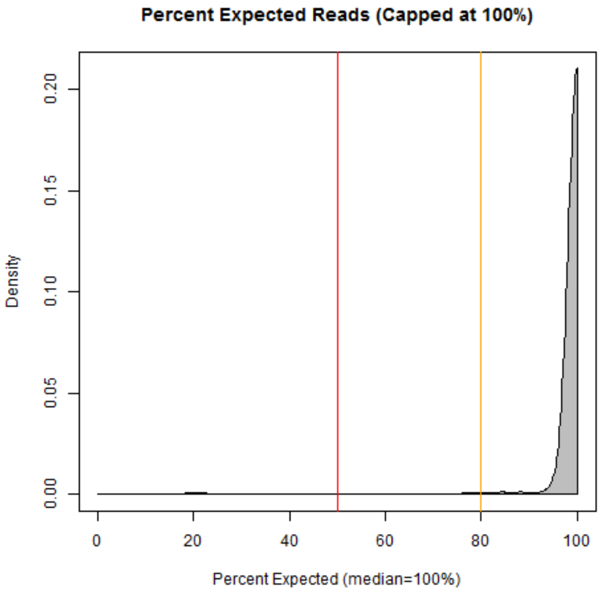
# Effect of Considering Only Samples Without Mixed Lanes

**(For *Entire* PE Run – otherwise, samples often combined between lanes)**

## Single-Barcode



## Dual-Barcode



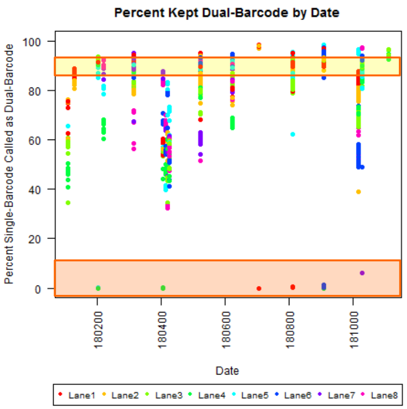
- By definition, these runs have a mix of barcode types; however, you can then check if all samples for a particular barcode type are within a subset of lanes (so, by lane within the run, the samples aren't mixed)
- Strictly speaking (separate barcode type for lane for entire run), this could be done for **3/18** runs for single-barcode samples
- This could be done for **2/18** runs for dual-barcode samples
- **While this noticeably reduces the number of runs to consider, it provides some evidence that avoiding mixed barcode designs may help reduce issues with having insufficient reads.**
  - There were also successful rapid runs with mixed barcode designs, but I would guess working with fewer samples at one time generally reduces the chance for human error.

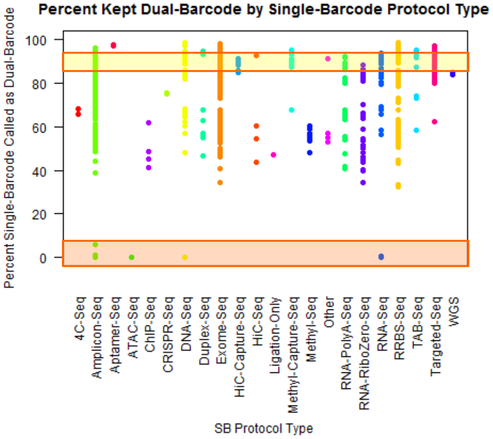
# Summary

- **We usually noticed a loss of reads per sample called with dual-barcodes versus single-barcodes (at least to some extent, as expected)**
- For SB as DB, sometimes the issue was unexpected 2nd barcodes
  - When fewer than ~5% of single-barcode samples were called with dual-barcodes, ~50% were **user-prepared** samples (where I didn't use the correct 2nd barcode, and that also wasn't in our records)
    - **My opinion is that user-prepared libraries should be separated by lane, so they can't affect other lab's samples**
- If those outliers are removed, there is a trend for average percent >Q30 (per lane, per run), which positively correlates to the SB-as-DB percent (*and samples with average >Q30 less than 80% on i7(I1) are more likely to have intermediate SingleBarcode-as-DualBarcode percentages*)
  - While number of runs is more limited, I2 Q30 seems to be more of an issue for *paired-end* libraries than *single-end* libraries (when multiplexed with a mix of single-barcode and dual-barcode samples).
- **So, I propose adding flags for returning base calling results if a) >40% of samples have <50% Q30 I2 Reads, b) if >50% of samples have <50% of the expected number of reads, and c) <200M PF Clusters per lane (for Regular HiSeq run).**
  - In this situation, I would recommend warning the user and considering re-doing the run if anybody reports anything strange about their results
- Whether this helps with goal of being able to avoid having to combine reads between runs (to predict/avoid read counts being too much lower than expected) is a little less clear, but I think there could be some benefit to understanding these results
  - For example, there is some tentative evidence that reducing the number of barcode types per run *may* help reduce the occurrence of not having sufficient reads (where I would guess human error may play *some* role).
  - This *may* indicate value in processing fewer samples at one time (with lower throughput Illumina sequencing), which could complement other indications that there may be value in having PI samples separated by lane (possibly prepared by the labs, to make sure full protocol is understood as clearly as possible)
  - For some applications, if you were concerned about the I2 quality scores, you could double the number of SE reads instead of using PE reads
    - If tested, perhaps having a longer single-barcode (I1) could be another solution?
- When I used the wrong second barcode, it was good that I could confirm the correct 2nd barcode. *However, this shows that up to 2% of (expected) reads mis-assigned to a sample (even when allowing 0 mismatches)*.
  - *True percentage of intended samples could be worse* (since the expected read count is often quite different between samples): hypothetically, assume that there is one ATAC-Seq sample that should have ~50,000,000 reads, and one Amplicon-Seq sample that should only have ~10,000 reads (0.02% the size of the ATAC-Seq sample). If some small percentage of reads are not assigned to the correct barcode, then I would expect that to be a bigger problem for the Amplicon-Seq sample (possibly having a larger percentage of ATAC-Seq reads incorrectly assigned to that sample).

# Additional Information

# Effect of Date and Number of Single-Barcode Samples on Percent Samples Kept (PE Runs)





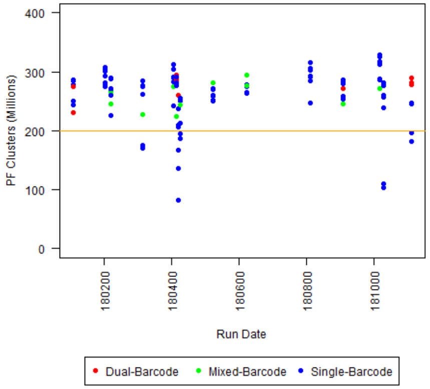
- Originally wanted to explain 5-15% loss (85-95% recovery), but lowest outliers were much lower
- **<2% Single-Barcode Kept as Dual-Barcode**
  - **User-Prepared Library**
    - Internal Scientist A (Amplicon-Seq + RNA-Seq): 180406 (TCTTTCCC → TATAGCCT), 180706 (TCTTTCCC → TATCCTCT or TATAGCCT), 180810 (TCTTTCCC → TATAGCCT)
      - One RNA-Seq sample mislabeled as DNA-Seq
    - Internal Scientist B (ATAC-Seq): 180810 (TCTTTCCC → GCGTAAGA), 180907 (TCTTTCCC → TATCCTCT or TATAGCCT), 180810 (TCTTTCCC → TATAGCCT)
    - Internal Scientist C (Amplicon-Seq): 180907 (TCTTTCCC → [not top])
  - **IGC Dual-Barcode Library (Called as Single-Barcode)**
    - External Scientist A (ATAC-Seq): 180203 (TCTTTCCC → CTCTCTAT, TCTTTCCC → TATCCTCT,
    - TCTTTCCC → CGTCTAAT, TCTTTCCC → TCTCTCCG)
    - Internal Scientist D (ATAC-Seq): 180406 (TCTTTCCC → CTCTCTAT)
    - Coloring checking for possible shifted / SNP mismatches, but Jinhui confirmed these were the expected 2nd barcodes
  - **So, these can all be explained by initially using the wrong 2nd barcode**

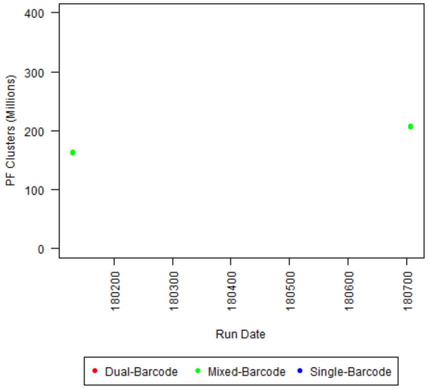
# Total Reads Per Lane (PE Runs)

## Regular Run

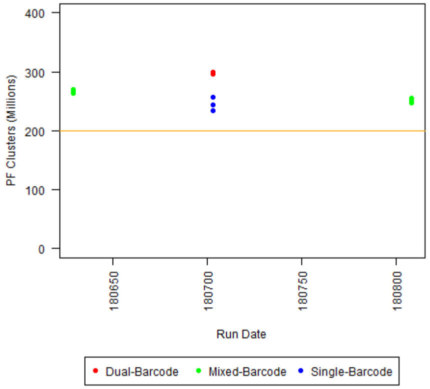
## Rapid Run

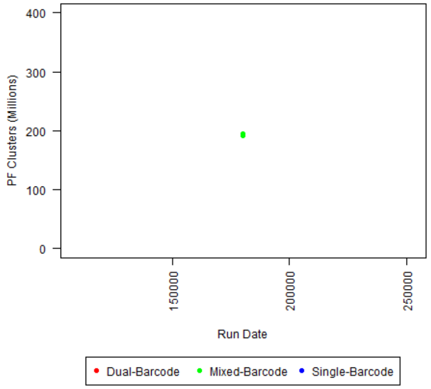
**Paired-End (PE)**





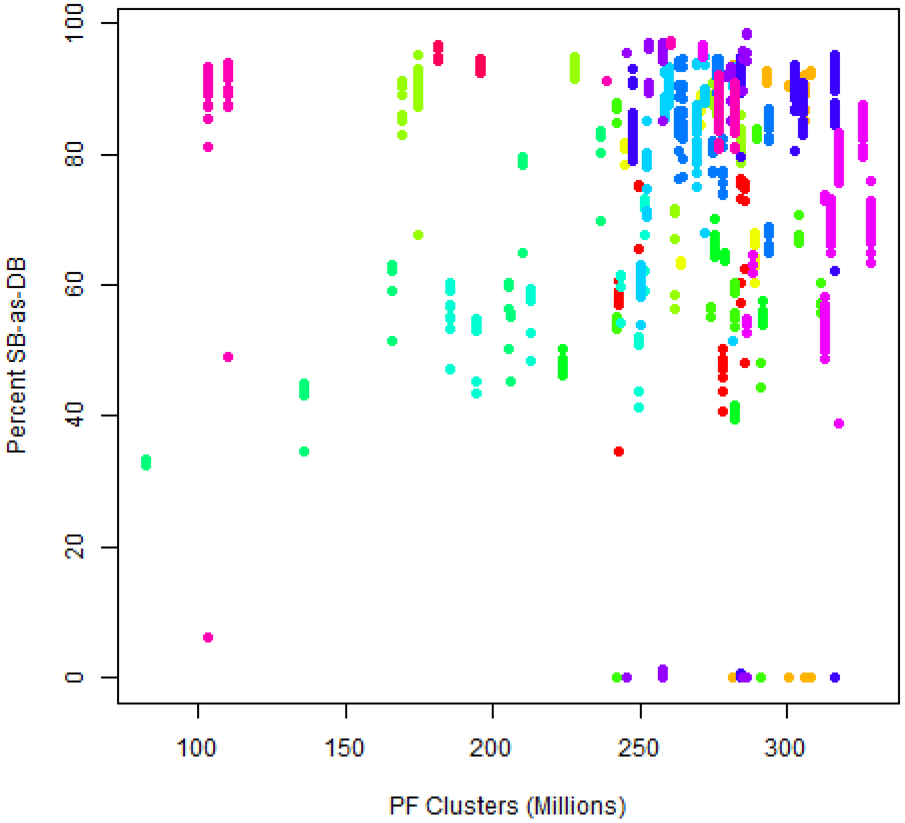
**Single-End (SE)**





- Do what extent does the total number of reads per lane affect the expect read percentages?
- **180524 and 181109 are the runs where <50% of SB PE samples have <50% of expected reads**
  - All lanes for 180524 had >200M reads (so, that was not among the runs with lanes that drop below 100M reads)
  - 2 lanes for 181109 have 180-200M reads (so, some decrease, but not the most extreme drop)
- **180406, 180907, and 181026 are the runs where <50% of DB PE samples have <50% of expected reads**
  - All lanes for 180406 and180907 had >200M reads
  - However,181026 does have 2 lanes with ~100M reads
- So, perhaps we should flag runs with less than 200M PF Clusters (for Regular Run), but that looks like a QC metric for a separate issue (than the worst loss of expected reads)

# SingleBarcode-as-DualBarcode Versus Total Reads (Regular HiSeq PE Runs Only)



- **We would also want to compare total reads to percent expected reads per sample, but those don’t separate cleanly by lane (we frequently combine reads between lanes in a run)**
  - Nevertheless, the total PF Clusters is not highly predictive of the percent SB-as-DB
  - *Possibly skip combing reads from abnormal lane?* You typically had more than one <200M lane per run (which you could tell from the previous slide)
    - So, you would usually be throwing out more than one lane, probably without precisely knowing the cause or if other samples in the run could be affected.
- Circumstantially, in terms of checking lanes with fewest PF clusters (regardless of whether reads were combined between lanes), the lane with the fewest total reads has a subset of RRBS samples (all single-barcode) for a project
  - The next lowest 2 lanes had a mix of barcode types (8+8 dual-barcode ATAC-Seq and 6-bp single-barcode Amplicon-Seq)

# Additional Single-End (SE) Run Notes

- Those are a little harder to find since single-barcode PE reads and dual-barcode PE reads each have 3 “reads” in basecalling folder (when visually inspecting folders to test additional basecalling). In contrast, the only way you can have 4 “reads” is if you have at least 1 dual-barcode PE sample in a run (or if you accidentally sequenced the extra index).
- Three runs had single-barcode miRNA-Seq samples that all had a different I2 (TCTTTCCC → GTTCAGAG; these would be like “wrong 2nd barcode” examples for PE runs)
  - One run also had “core test” miRNA-Seq samples with varying 2nd barcodes (but one had GTTCAGAG as the 2nd barcode matching the most reads).
  - Likewise, there was a “smarter” stranded RNA-Seq library with a different I2 in another run (TCTTTCCC → CCTATCCT). This was considered a dual-barcode sample (rather than a single-barcode sample) in an earlier mixed run (one of which needed additional reads).

# Unassigned Reads with Combined Sample Sheet
## (Dual-Barcode Samples with SingleBarcode-as-DualBarcode Samples)

| Lane | Total Reads | Unassigned Reads |
|---|---|---|
| 1 | 253,035,514 | 17,419,558 **(6.9%)** |
| 2 | 251,599,358 | 14,052,646 **(5.6%)** |
| 3 | 249,597,424 | 15,564,126 **(6.2%)** |
| 4 | 248,589,749 | 14,224,124 **(5.7%)** |
| 5 | 247,398,933 | 14,812,576 **(6.0%)** |
| 6 | 249,217,830 | 14,710,545 **(5.9%)** |
| 7 | 252,138,822 | 13,825,848 **(5.5%)** |
| 8 | 254,409,849 | 14,904,837 **(5.9%)** |

- Most results only look at the **single-barcode** samples converted to dual-barcode samples
  - So, this is the percent of unassigned reads when single-barcode samples are ***treated like dual-barcode samples***
- However, as requested by Xiwei, I created a reformatted ***combined*** sample sheet for this selected run (with SingleBarcode-as-DualBarcode samples, *in addition to* dual-barcode samples)
  - In this particular run, all dual-barcode samples were miRNA-Seq and all single-barcode samples were ChIP-Seq
  - So, I would say the unassigned read rate (for SE Dual-Barcode reads) was reasonable (5.5-6.9%)

# Unassigned Reads with Combined Sample Sheet
## (Use Representative PE Samples with Low SB-as-DB from Slide #10)

### 180413_B

| Lane | Total Reads | Unassigned Reads |
|---|---|---|
| 1 | 286,017,403 | 44,993,213<br>(15.7%) |
| 2 | ?? | NA |
| 3 | 294,570,287 | 33,399,207<br>(11.3%) |
| 4 | 223,451,103 | 109,055,604<br>(48.8%) |
| 5 | 282,118,324 | 169,585,369<br>(60.1%) |
| 6 | 291,824,958 | 139,083,251<br>(47.7%) |
| 7 | 279,053,751 | 104,675,922<br>(37.5%) |
| 8 | 275,711,708 | 97,694,679<br>(35.4%) |

### 180420_A

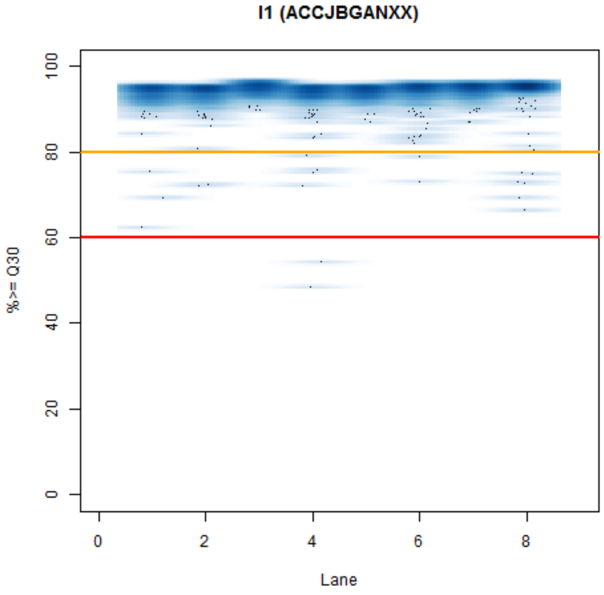
| Lane | Total Reads | Unassigned Reads |
|---|---|---|
| 1 | 259,647,257 | 21,795,036<br>(8.4%) |
| 2 | 166,015,607 | 75,433,188<br>(45.4%) |
| 3 | 135,569,697 | 87,260,176<br>(64.4%) |
| 4 | 205,546,481 | 87,260,176<br>(42.4%) |
| 5 | 236,695,633 | 53,363,042<br>(22.5%) |
| 6 | 209,811,004 | 58,404,804<br>(27.8%) |
| 7 | 205,976,230 | 102,588,814<br>(49.8%) |
| 8 | 82,023,712 | 63,803,201<br>(77.8%) |

### 181015_A

| Lane | Total Reads | Unassigned Reads |
|---|---|---|
| 1 | 326,062,934 | 58,348,140<br>(17.9%) |
| 2 | 317,715,697 | 73,102,859<br>(23.0%) |
| 3 | 328,458,109 | 106,876,576<br>(32.5%) |
| 4 | 315,048,437 | 100,925,785<br>(32.0%) |
| 5 | 271,300,439 | 33,955,849<br>(12.5%) |
| 6 | 312,817,879 | 111,780,187<br>(35.7%) |
| 7 | 286,647,318 | 272,492,204<br>(95.1%) |
| 8 | 288,338,385 | 271,705,314<br>(94.2%) |

- I1 for single-barcode was 6 bp; I1 for dual-barcode was 8bp (8+8 design)
  - These are unassigned reads when treated like **dual-barcodes**. Because these samples also have **low SingleBarcode-as-DualBarcode percentages**, the *unassigned* reads would be *lower* if treated as **single-barcode** samples. So, the read loss was less of an issue when only using the 1st barcode (which is what was returned for single-barcode samples).
- Highlight lanes that have mixed single-barcode and dual-barcode samples
  - One run was mixed overall but had barcode-types were isolated per-lane (so, dual-barcode only lanes were highlighted)
  - All other lanes (not highlighted) are all single-barcode
  - Even without mixing, lower Q30 runs have more unassigned reads. Matching plot on slide #10, percent aligned reads is better on 1st lane of middle run (but still problematic overall, and a little worse than previous slide)

# Example of 2019 SE Regular Run with Only Single-Barcode

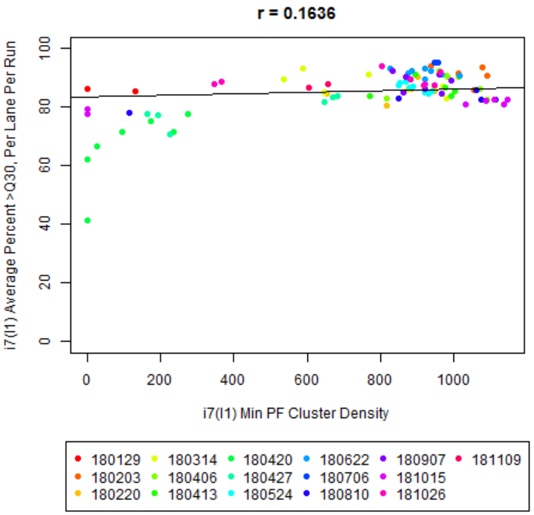


| Lane | Total Reads | Unassigned Reads |
|---|---|---|
| 1 | 279,287,836 | 10,191,359<br>**(3.6%)** |
| 2 | 277,738,123 | 10,331,191<br>**(3.7%)** |
| 3 | 241,694,667 | 6,689,740<br>**(2.8%)** |
| 4 | 272,853,908 | 9,609,021<br>**(3.5%)** |
| 5 | 272,715,683 | 7,816,687<br>**(2.9%)** |
| 6 | 258,714,003 | 10,185,119<br>**(3.9%)** |
| 7 | 259,880,906 | 9,041,522<br>**(3.5%)** |
| 8 | 257,000,873 | 9,598,470<br>**(3.7%)** |

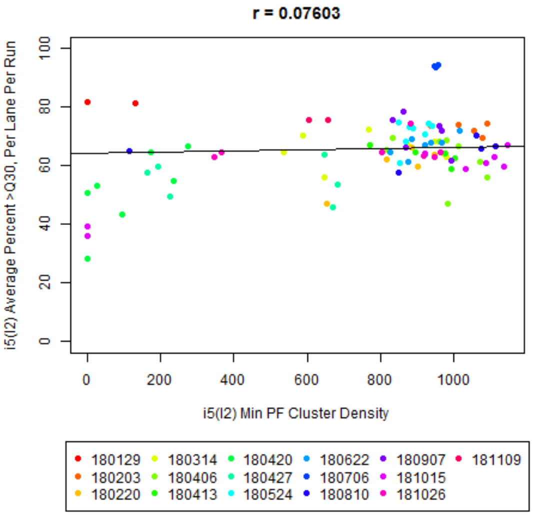
- Single-barcode single-end unassigned read rate is still ~50% (**3% versus 6%**) that of the mixed dual-barcode on slide #26 (even though the results on slide #26 are way better than slide #27)

# Comparing I2 Percent >Q30 with Minimum PF Cluster Density

## I1



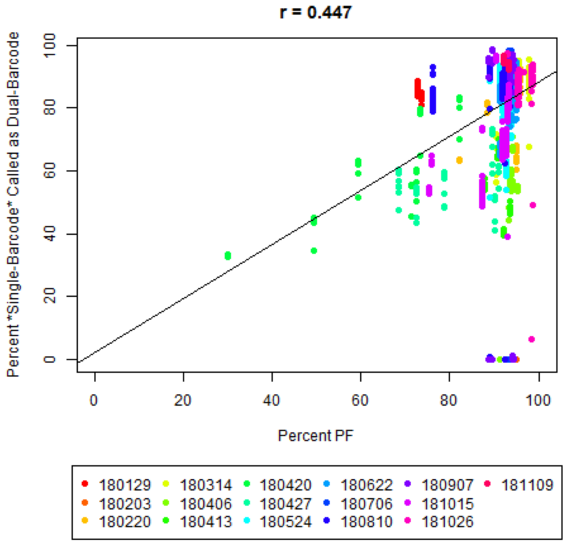
## I2



- This shows that the samples with the lowest SingleBarcode-as-DualBarcode rate also has tiles with a PF density of 0
- While I2 Percent >Q30 less than 60% probably sufficient for warning to look closer, perhaps I2 Percent >Q30 percent less than 50% or 40% is better as a threshold seriously consider re-processing the samples and not using the current run
- The trend is more clear for I1 than I2, probably because the PF clusters are defined at an earlier cycle
  - Minimum PF Cluster density is the <u>same</u> for I1 and I2, but Percent >Q30 is <u>different</u> for I1 and I2 (so, the trends above can be different)

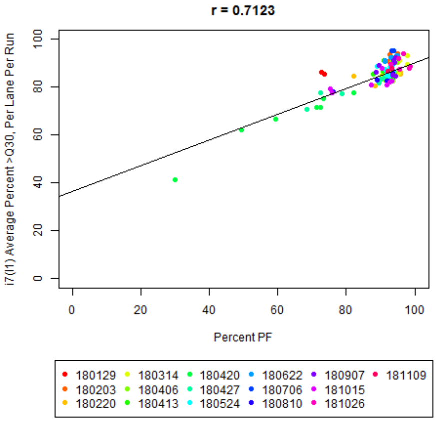
# Comparisons with Percent Passing Filter (PF)

## SingleBarcode-as-DualBarcode

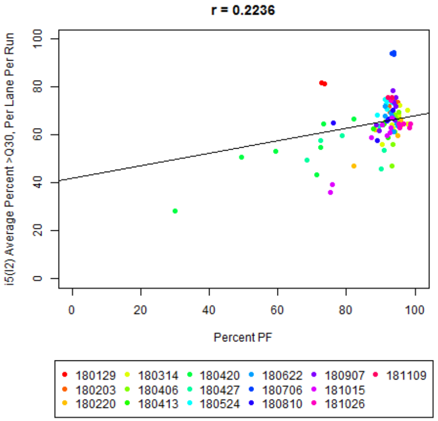


## Index Q30

### I1



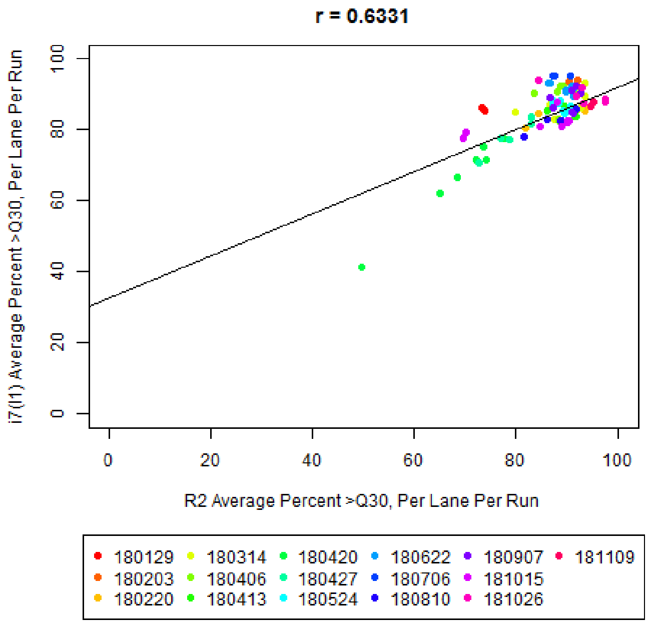
### I2



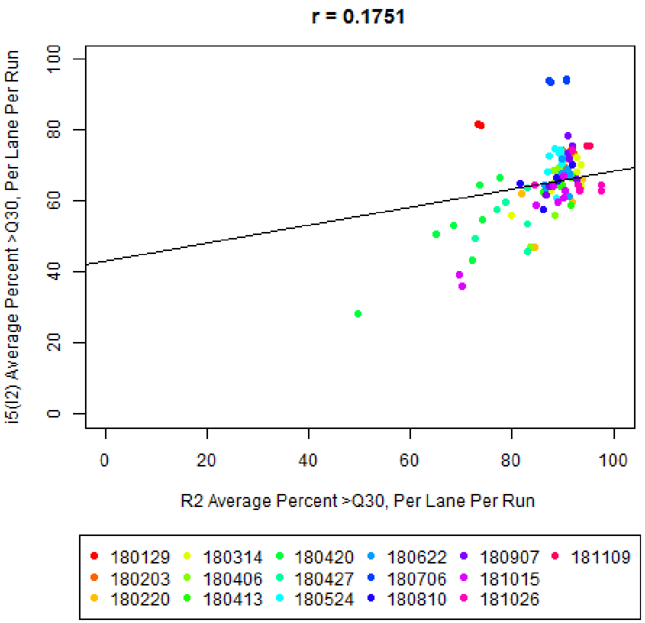
- Slightly worse correlation with SingleBarcode-as-DualBarcode (relative to I2 Q30, **slide #8**), but stronger association with I1 Percent >Q30 than PF Cluster Density (previous slide)

# Correlation Between Percent >Q30 for I2 versus R2 (Mixed Paired End Runs)

## I1



## I2



- If I2 quality scores drop, you may also expect R2 quality scores to drop
- As expected, average I2 quality scores are lower than average I1 quality scores (just like R2 quality scores are lower than R1 quality scores)
- However, the I1 quality score is more highly correlated with the R2 quality score than I2 (and it is physically closer).