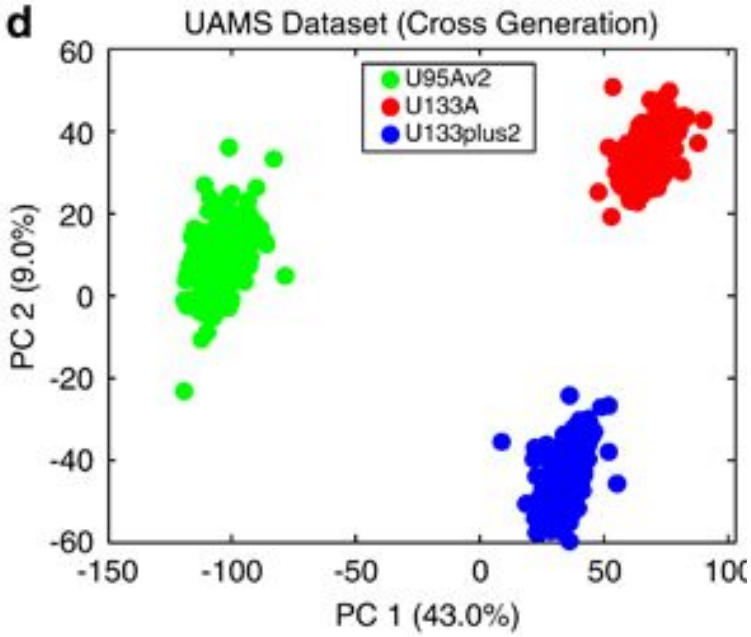
NovaSeq 6000 vs HiSeq 2500
Evaluation by mRNA-seq

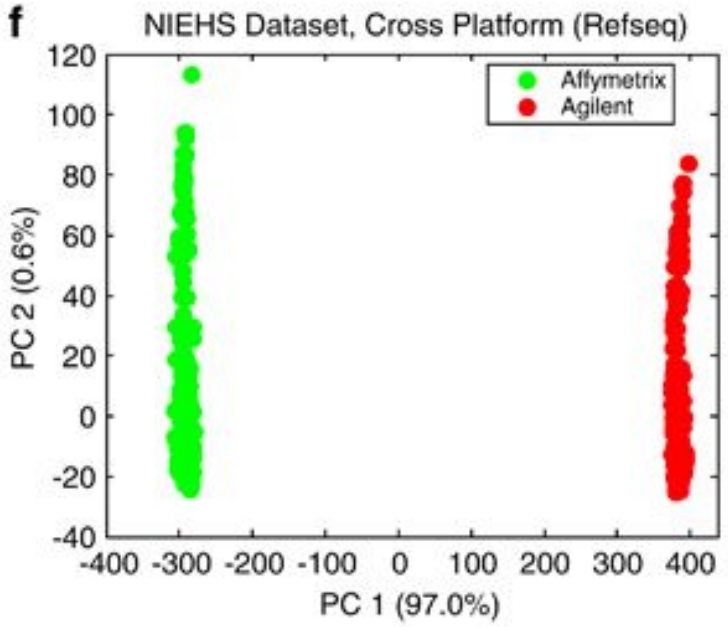
Joel Parker

Lineberger Bioinformatics Core

https://lbc.unc.edu/

Instrument Bias



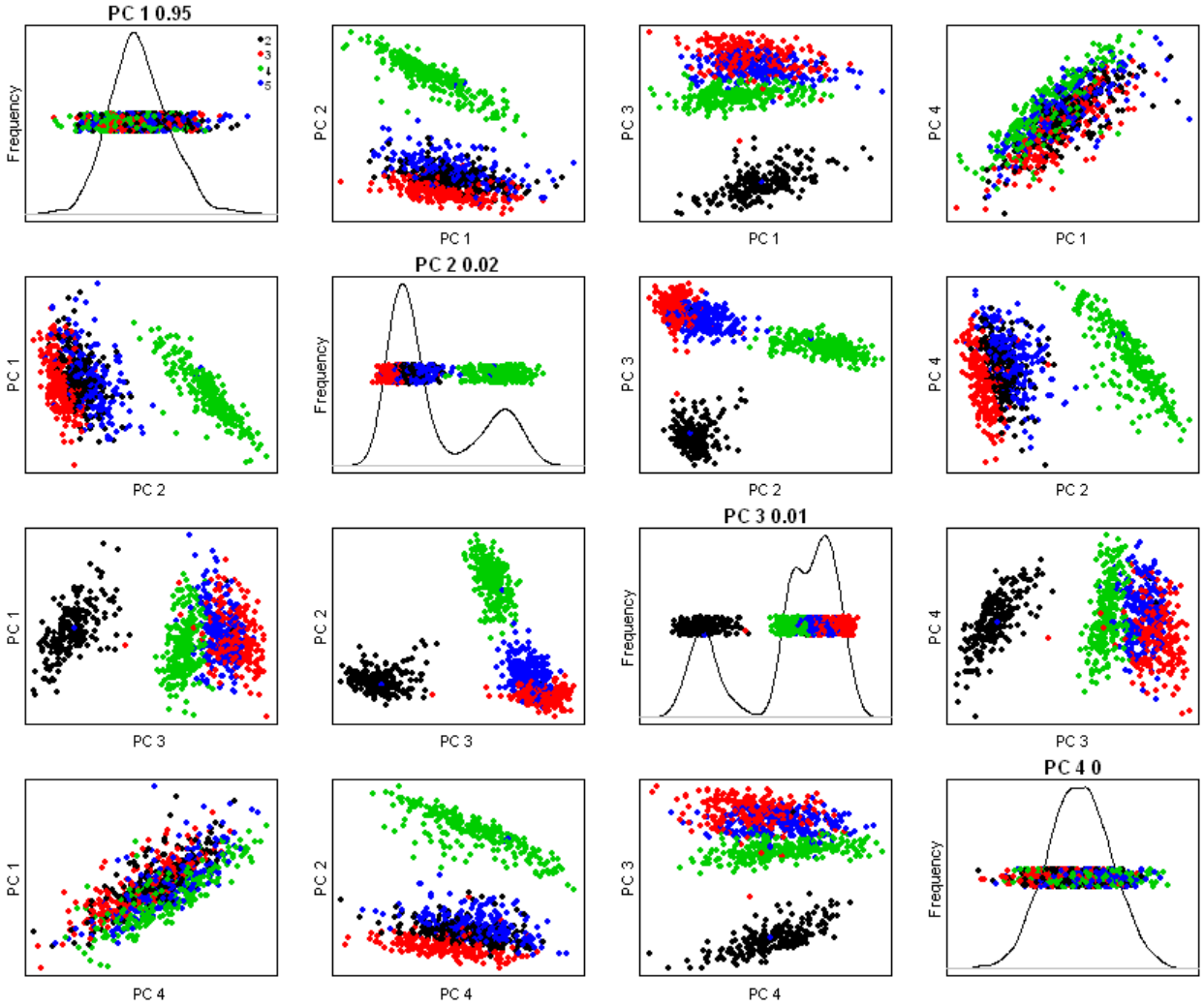


Platform	Samples in GEO
U133plus2	146,142
U133A	22,283
U95	6,446
Agilent 4x44k	14,367

Luo et al., MAQC-II 2010

UNC LINEBERGER

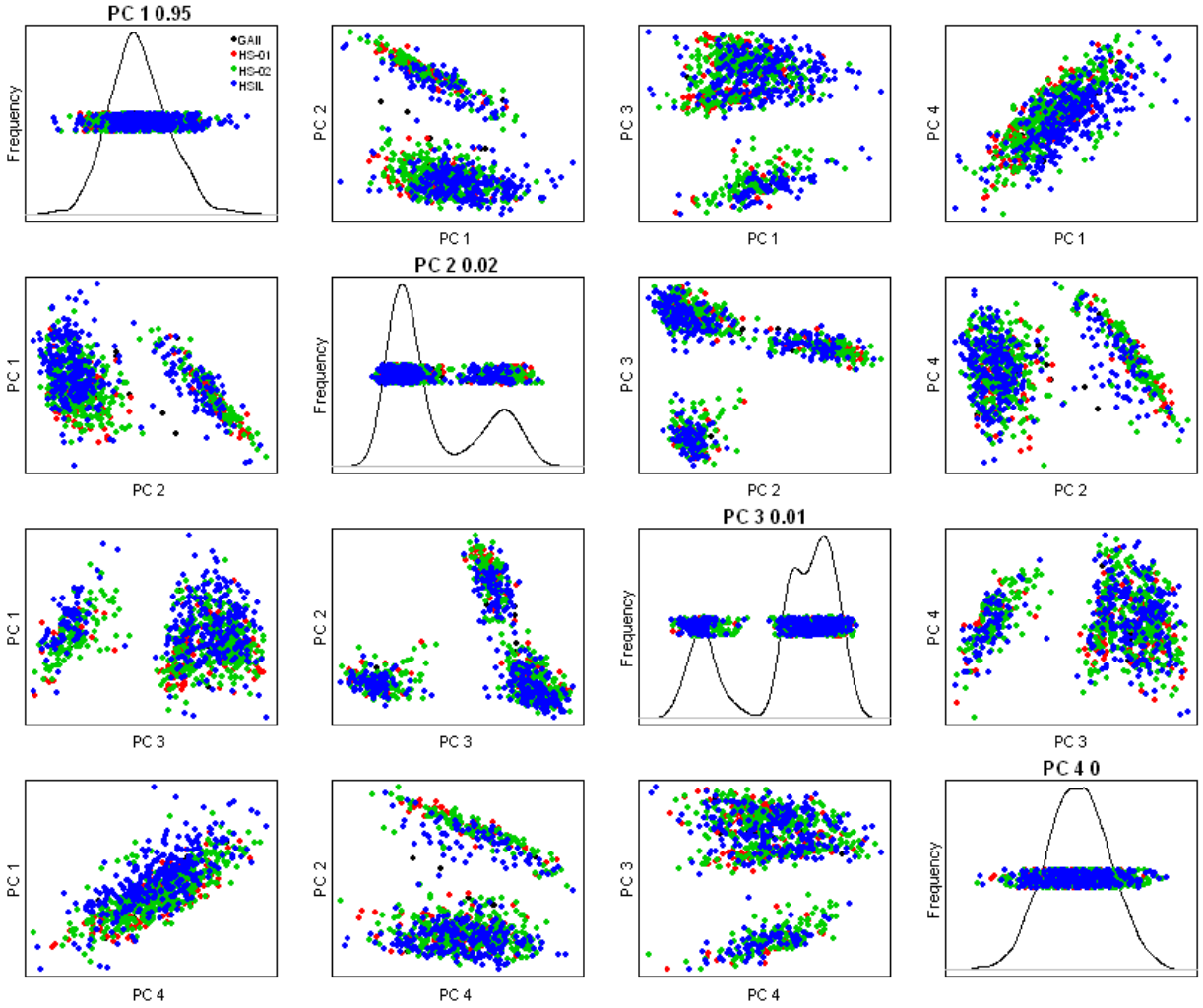
Instrument bias in mRNA-seq



CD19
CD8
CD14
CD4

UNC LINEBERGER

Instrument bias in mRNA-seq



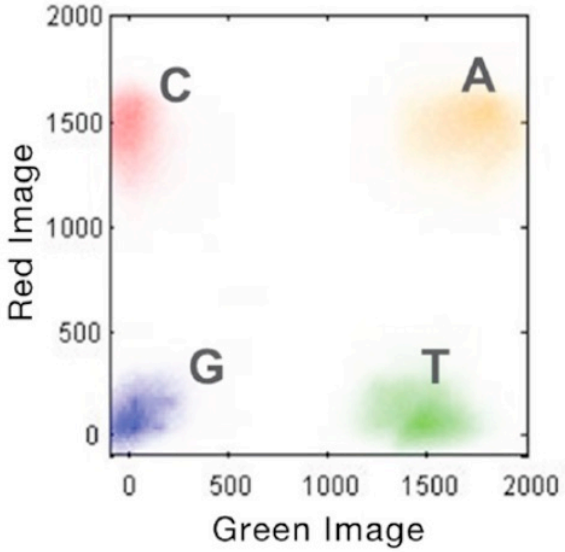
4 machines
2 sites
6 batches – 2 mos

GAIIx
HS-01
HS-02
HS-IL

UNC LINEBERGER

Hiseq vs Novaseq

- Hiseq2500 uses 4 channels while Novaseq only uses 2
  - G is represented by the lack of signal, previously called N
  - Poor quality reads may show up as polyG

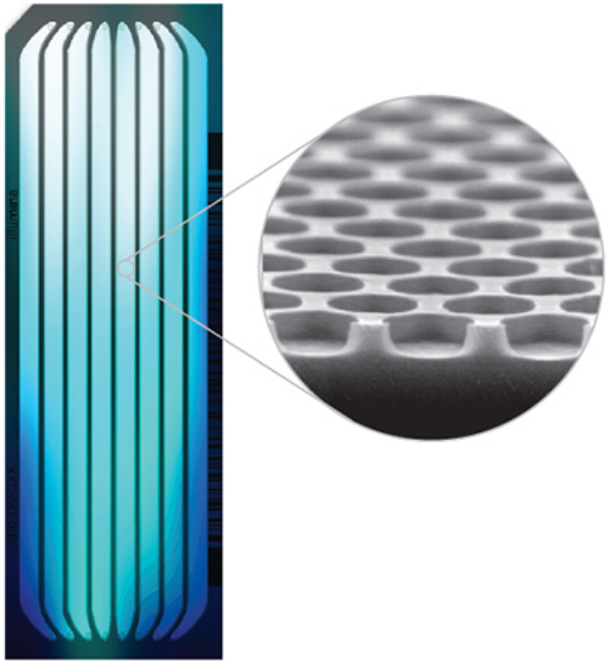


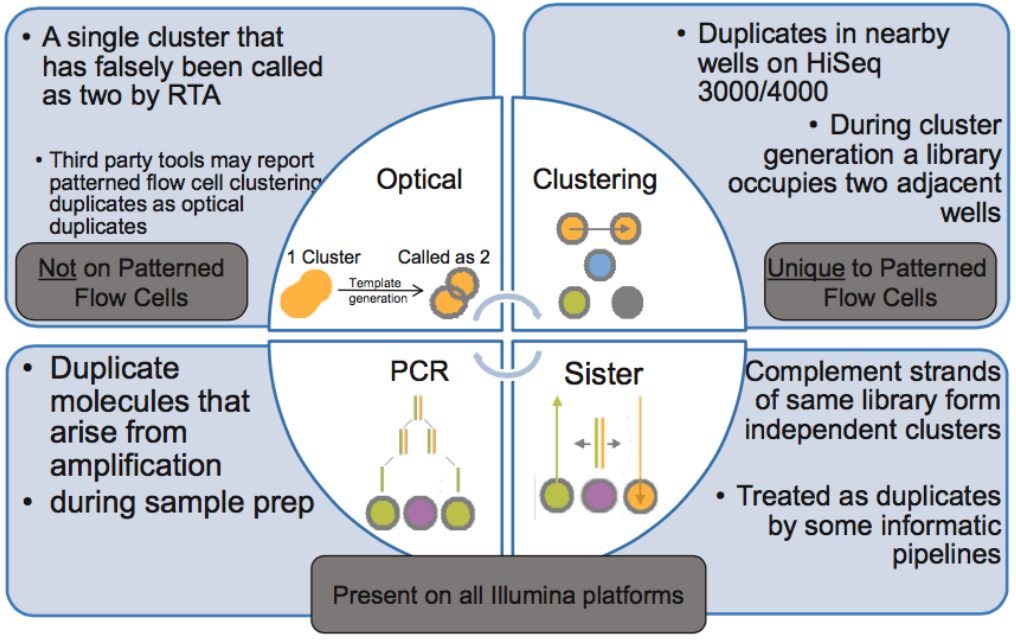
- Increased sensitivity to imbalanced indexes during calibration
- cutadapt permits removing trailing ‘G’s

https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology/2-channel-sbs.html

Hiseq vs Novaseq

- Ordered flowcells
  - Reports of barcoded reads spilling over into adjacent wells



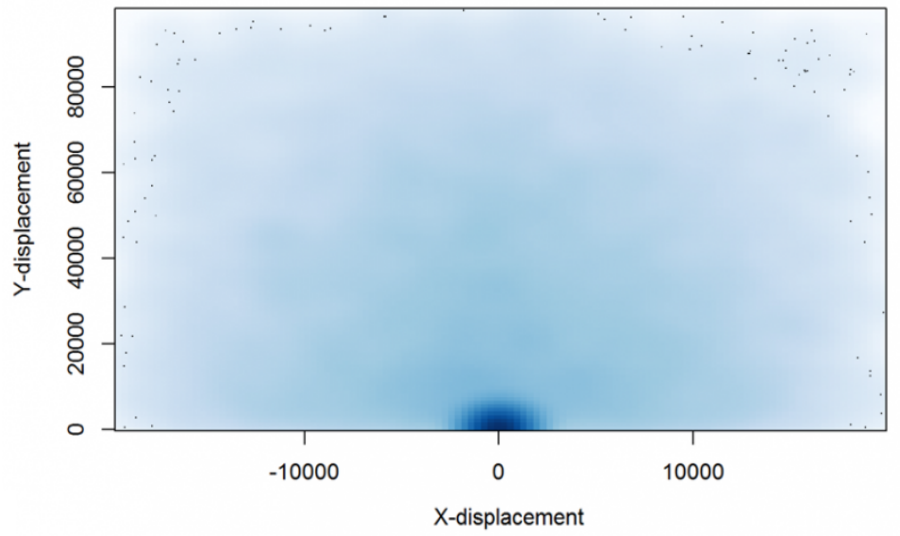


Hiseq vs Novaseq

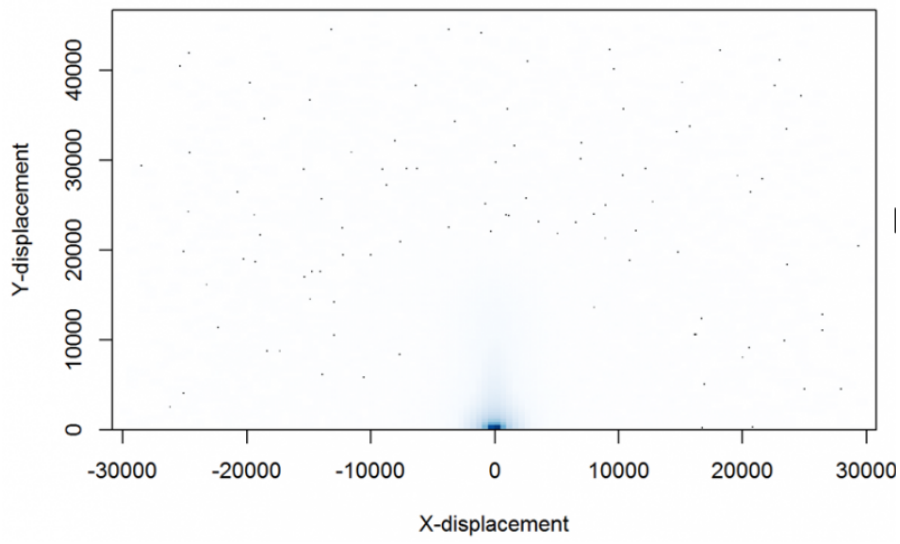
- Ordered flowcells
  - Reports of barcoded reads spilling over into adjacent wells

Relative position of duplicates

Hiseq 2500



Novaseq



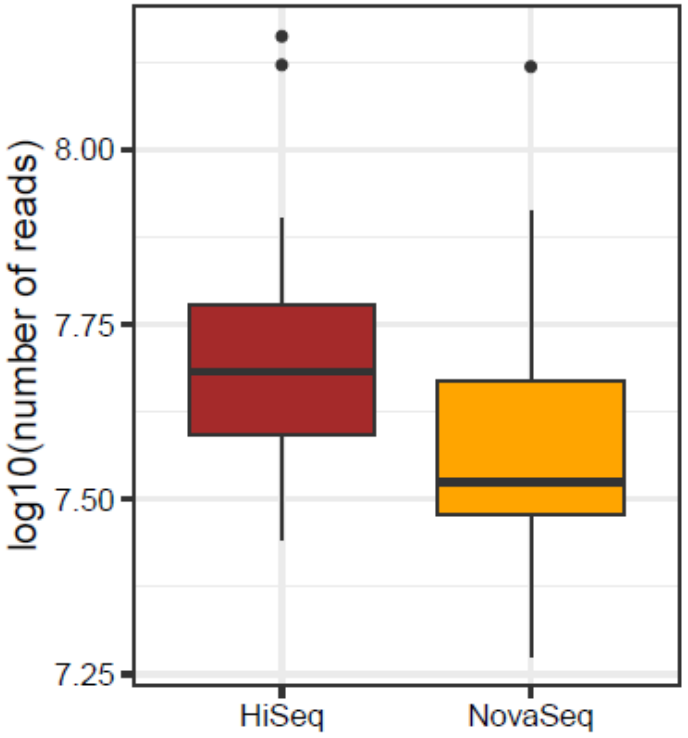
https://sequencing.qcfail.com/articles/illumina-patterned-flow-cells-generate-duplicated-sequences/

UNC LINEBERGER

Evaluation

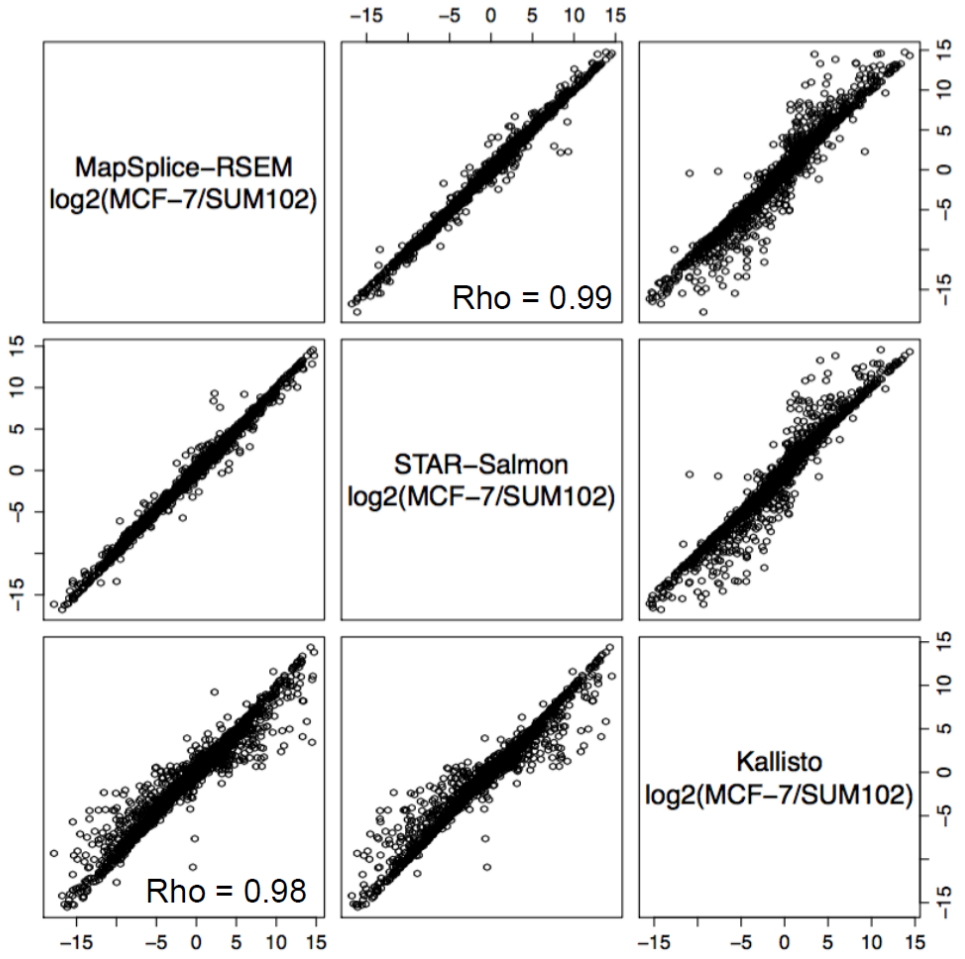
- RNA isolated from 48 GEMM derived tumors
- TruSeq stranded mRNA
- Single barcode
- 2 NovaSeq lanes (S1) with 24 samples / lane
- 12 HiSeq lanes with 4 samples / lane

- Identical library on both machines

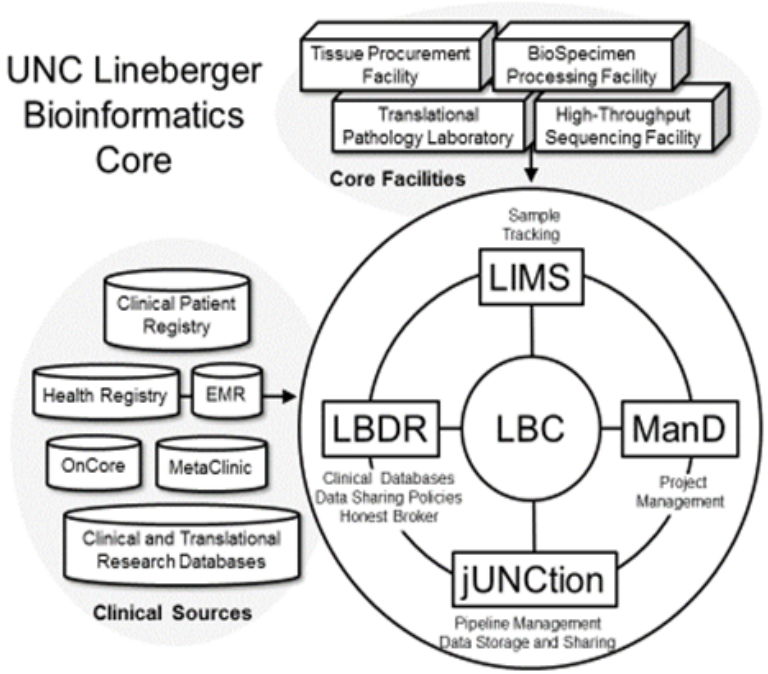


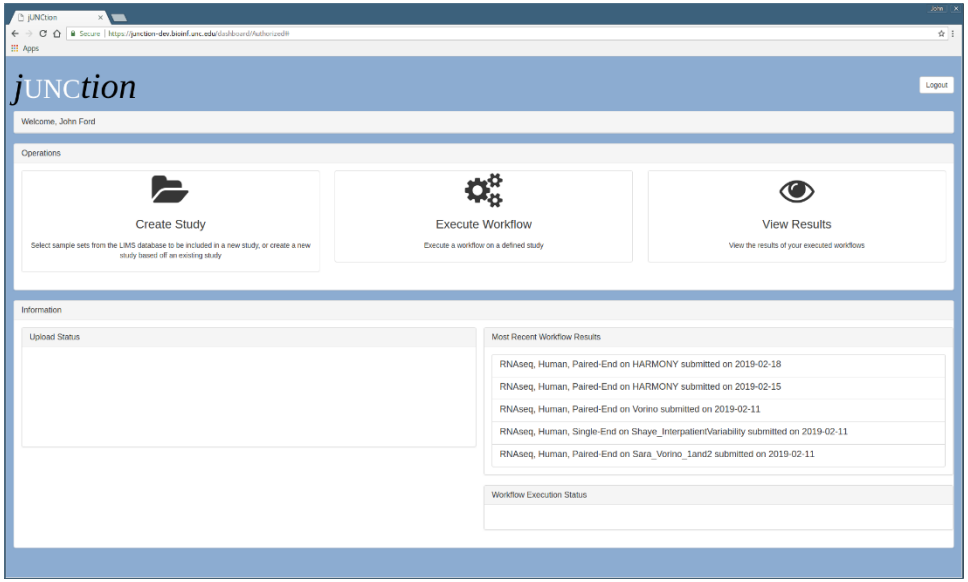
RNA-seq Workflow

	Legacy	Current
Alignment	MapSplice	STAR
Expression	RSEM	Salmon
Normalization	Upper quartile	Upper quartile
Genome	GRCh37	GRCh38
Transcriptome	UCSC knownGene	GenCode v22
Runtime (min)	~800	80
GB RAM	16	4



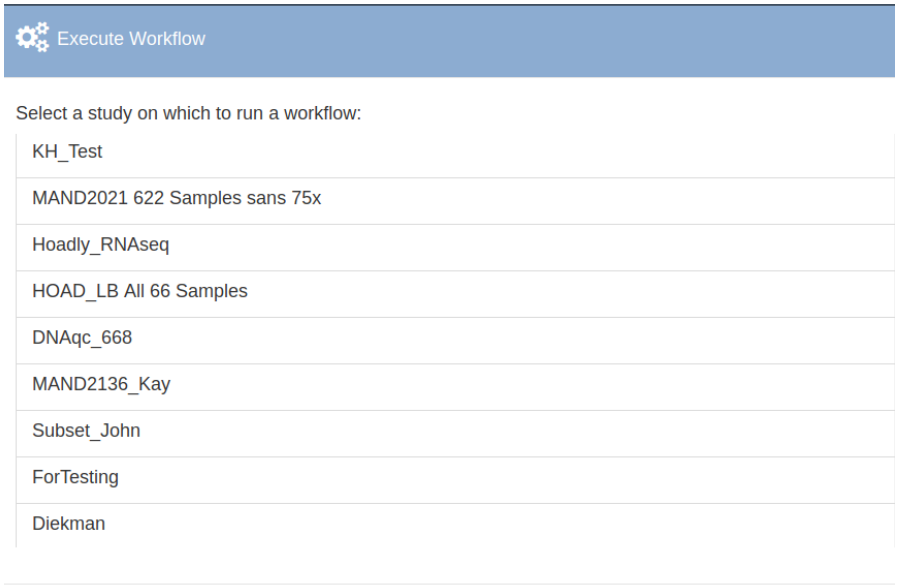
RNA-seq Workflow





UNC LINEBERGER

RNA-seq Workflow



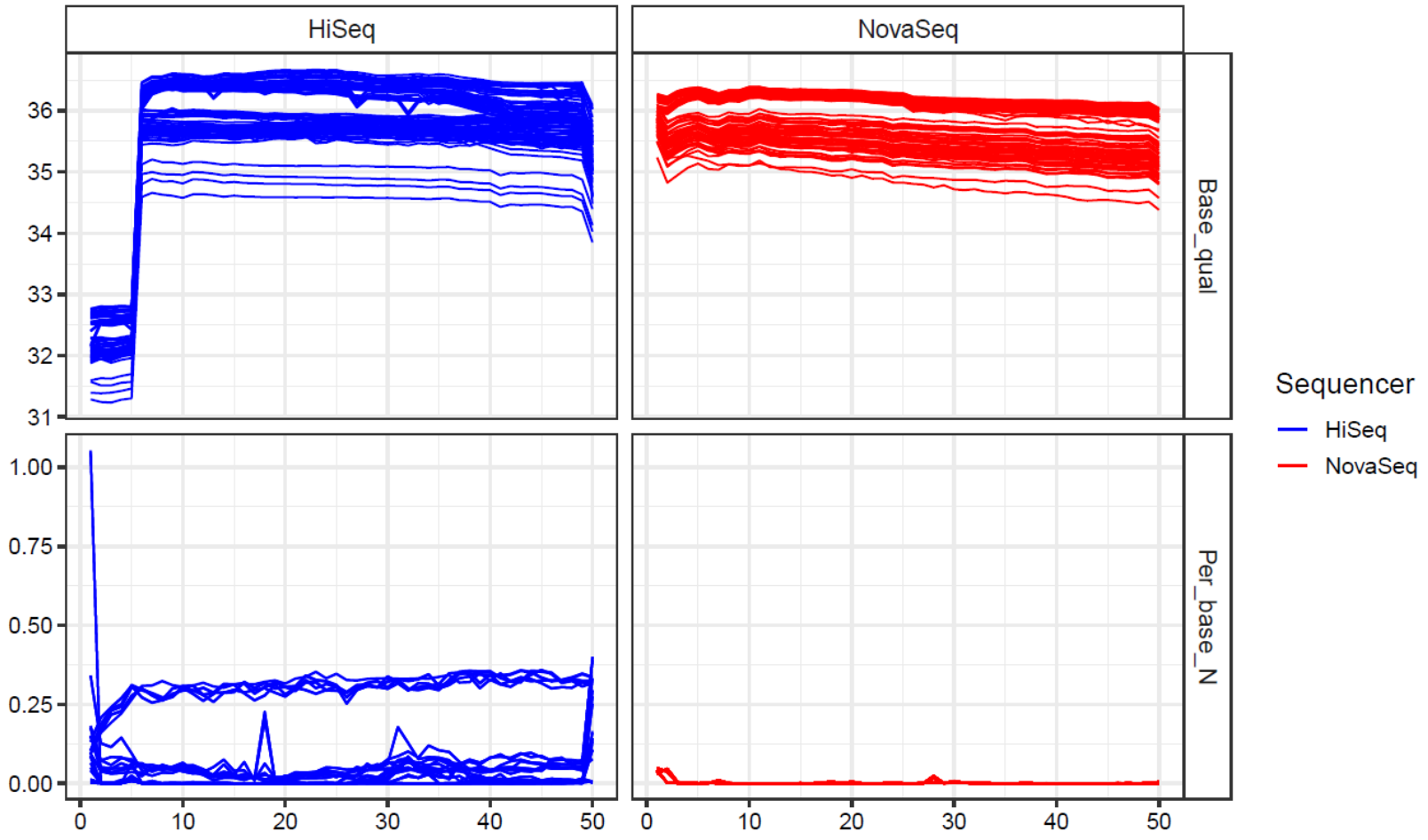
User & Lab (group) level security policies based on onyen

Provenance

User friendly

UNC LINEBERGER

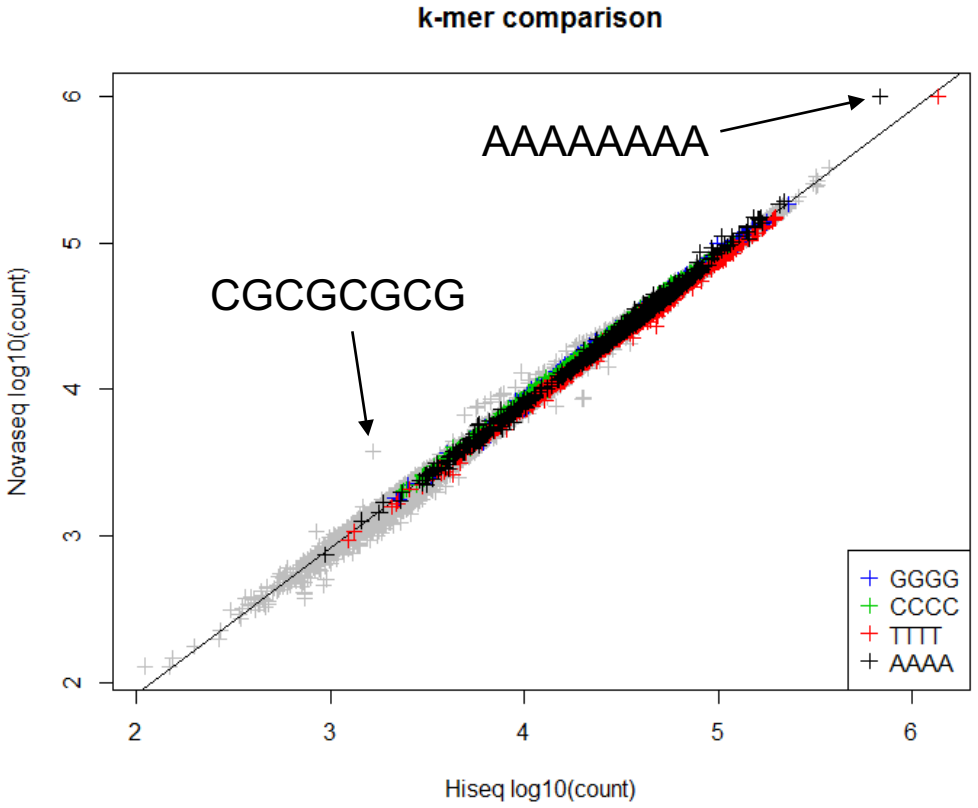
Base Quality



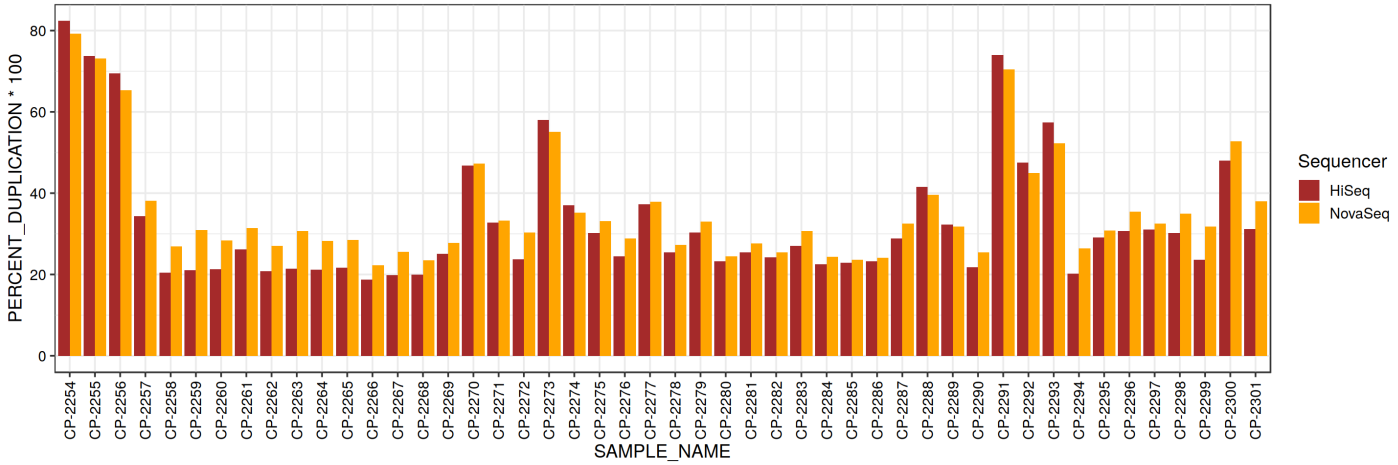
Novaseq base qualities are binned to
2, 12, 23, 37

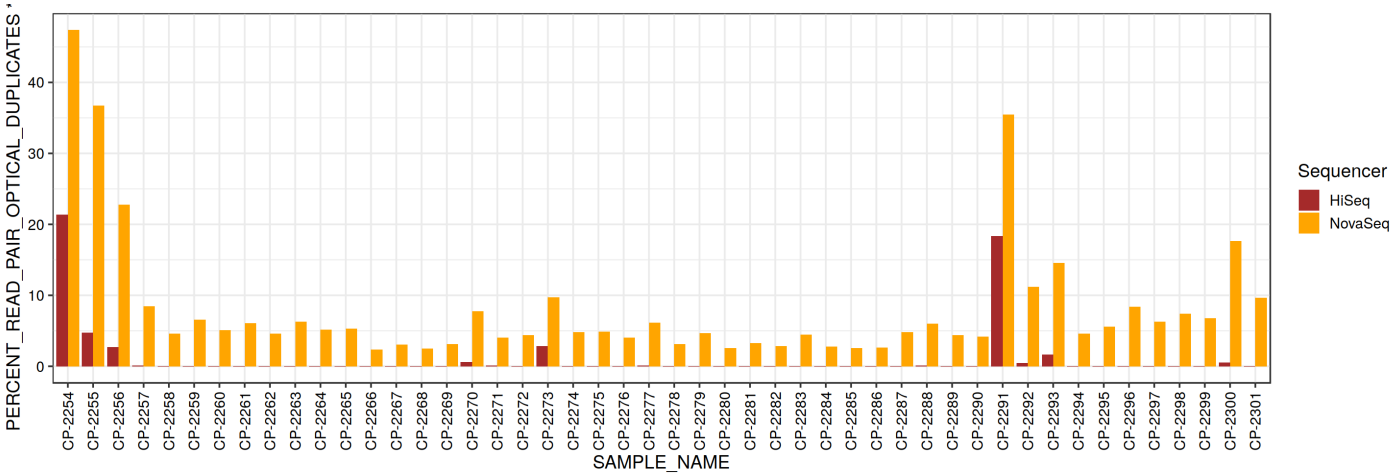
Base composition

- K-mer counts (k=8 shown) tabulated from fastqs
- Expected bias in G rich sequences is not observed
- polyA and CG repeating sequences are mildly enriched in the Novaseq run



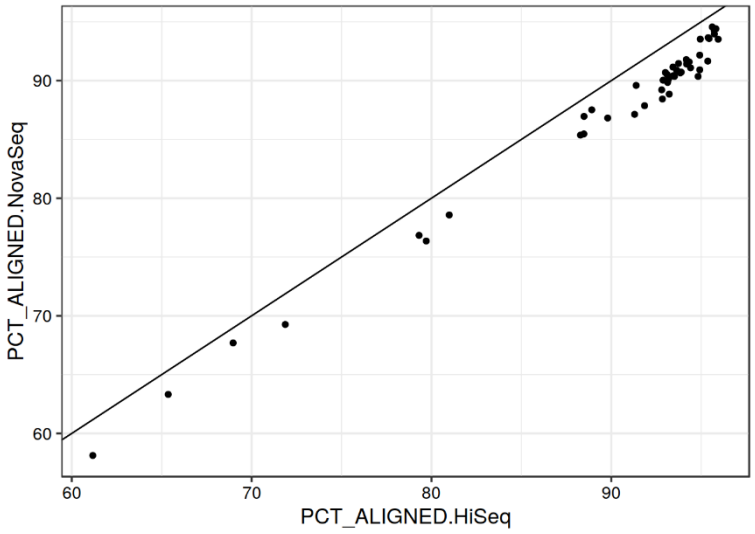
Duplication

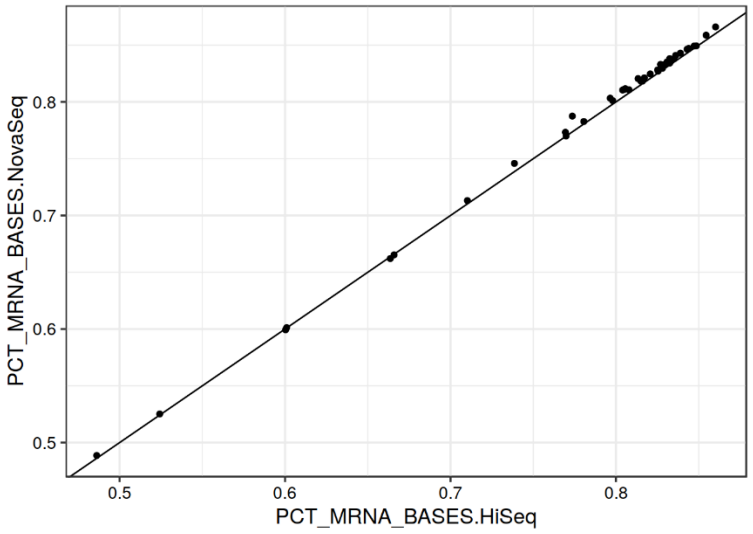


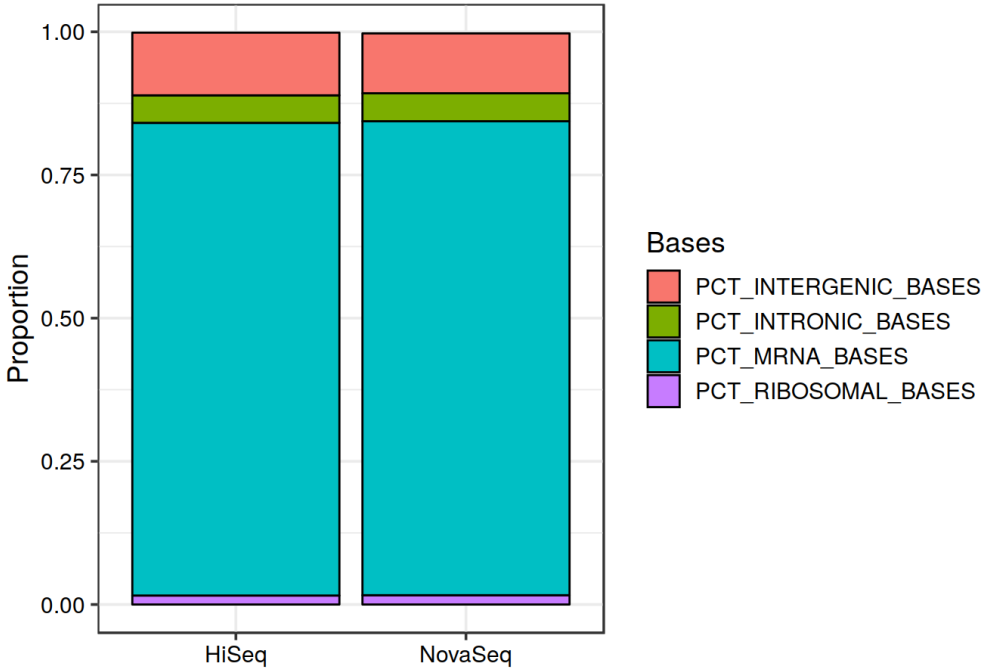


UNC LINEBERGER

Alignment

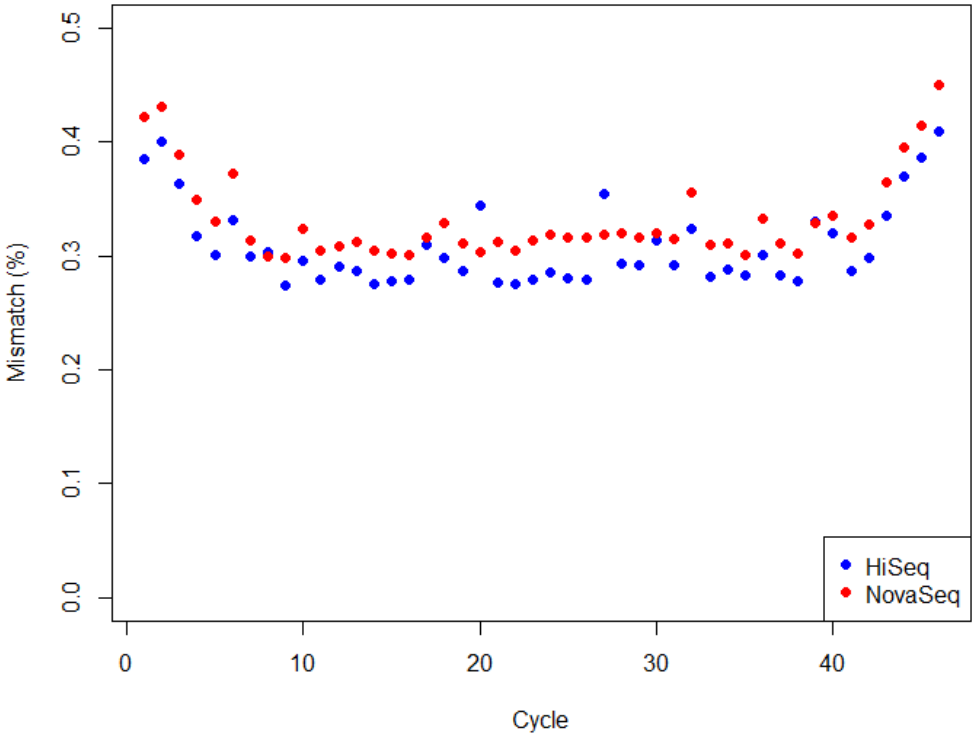


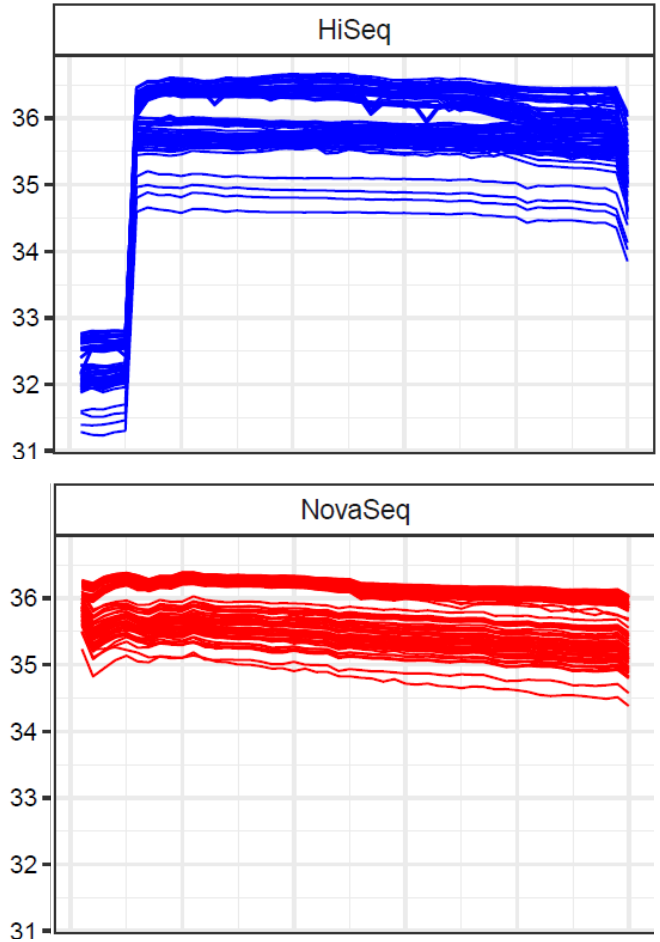




UNC LINEBERGER

Alignment



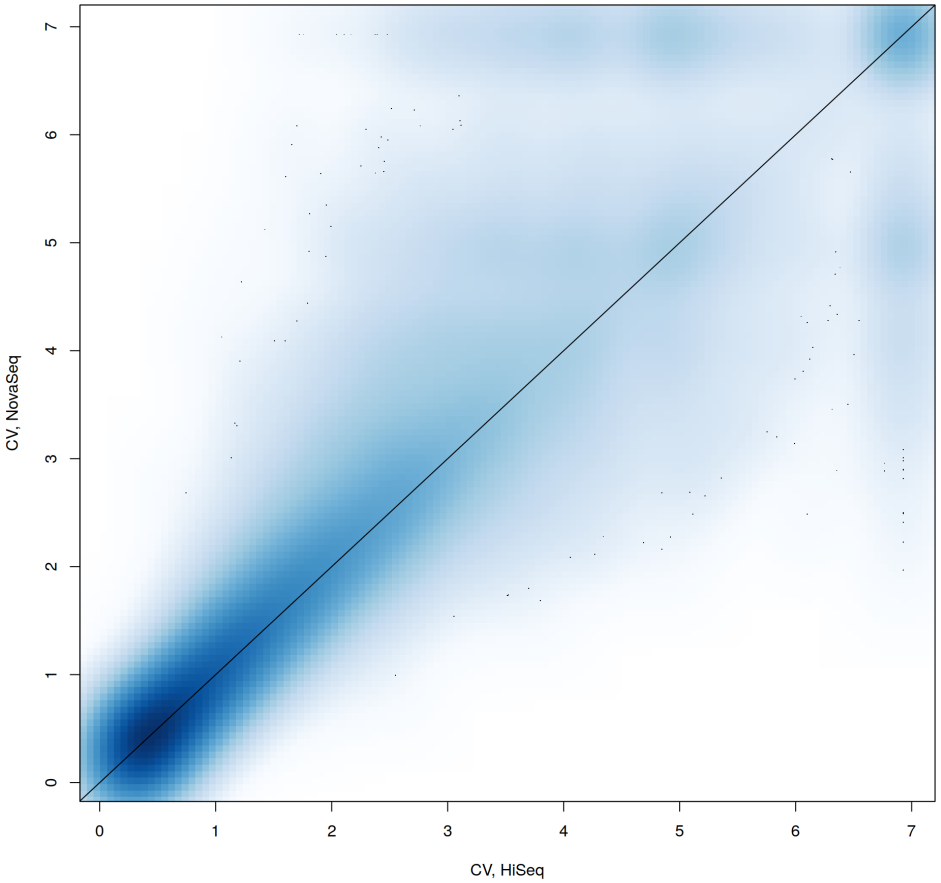


- Modest increase in mismatch rate
- Estimated Q scores do not appear accurate in the initial cycles

http://lh3.github.io/2017/07/24/on-nonvaseq-base-quality

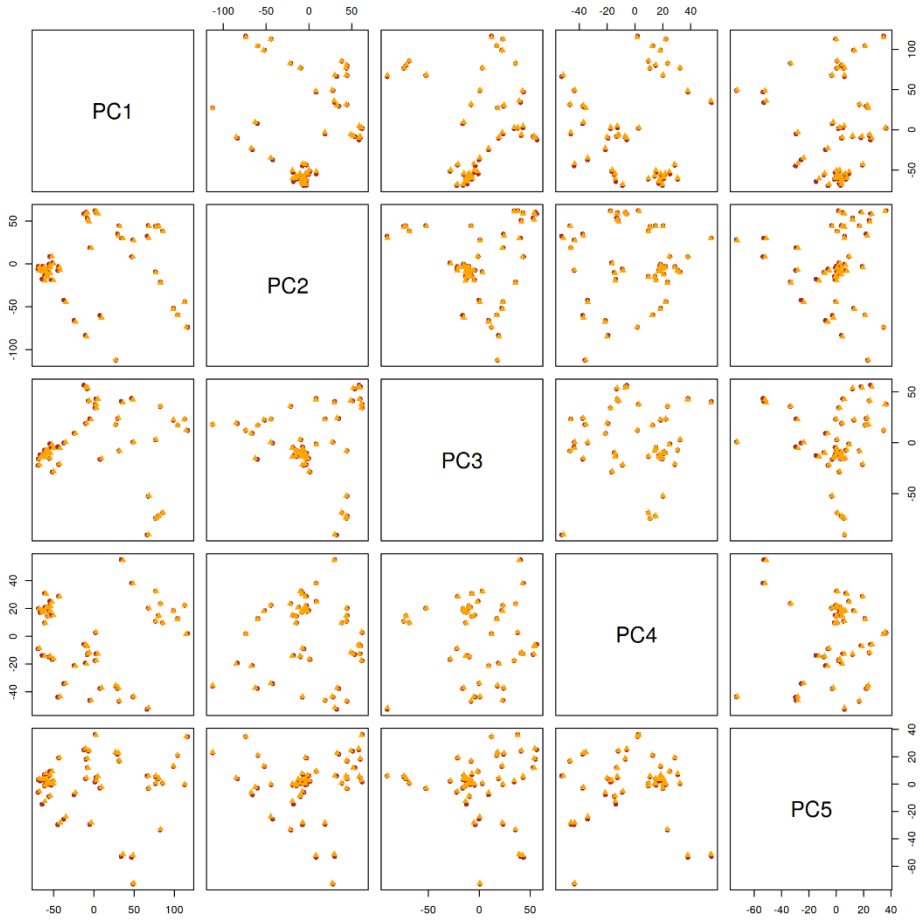
Repeatability

- Gene coefficient of variations (CV) estimated for each cohort
- CV estimates indicate high concordance in expression variation



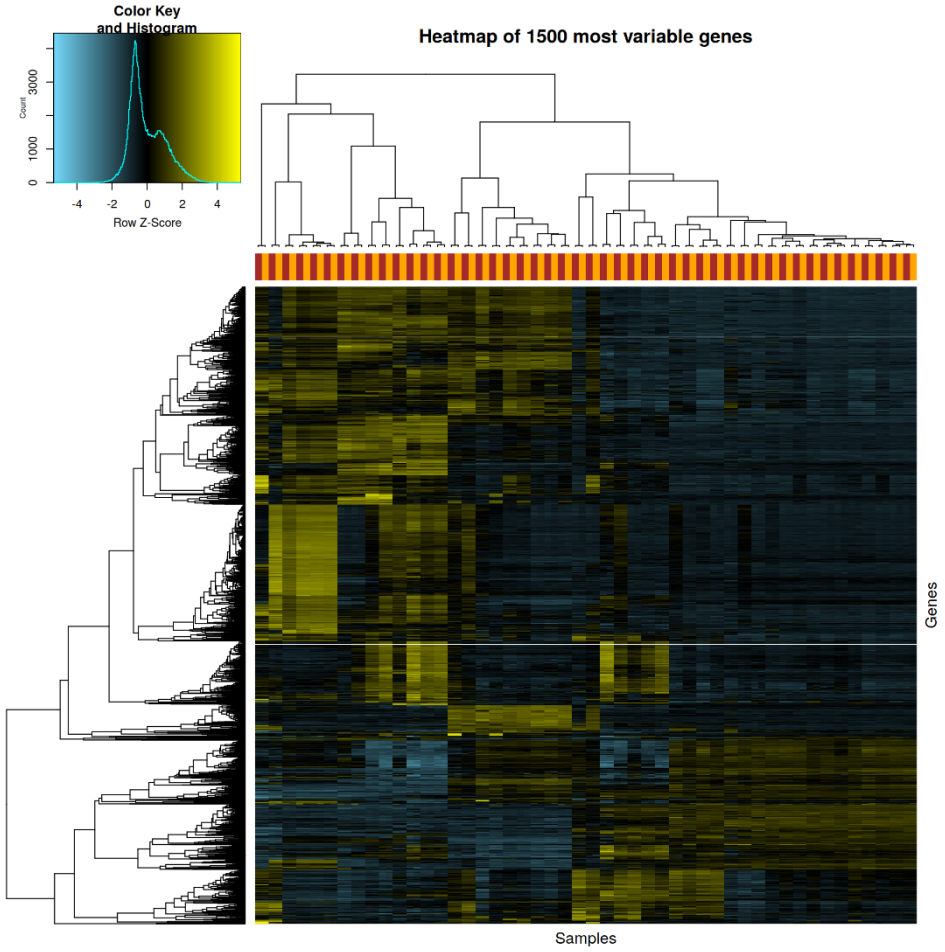
UNC LINEBERGER

Unsupervised Comparison



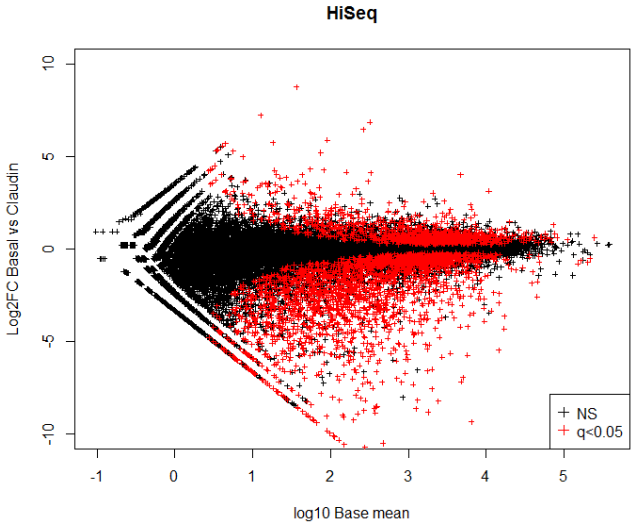
- PCA performed on log transformed, upper quartile normalized count estimates
- Nuisance variation due to instrument bias is not observed in any of the top components of variation
- PC35 (0.62% of expression variation) is the most highly ranked component associated with instrument (WRS $p < 0.001$)

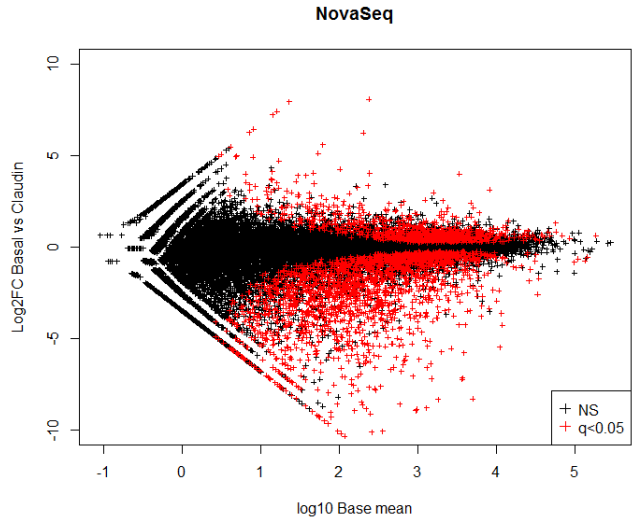
Unsupervised Comparison



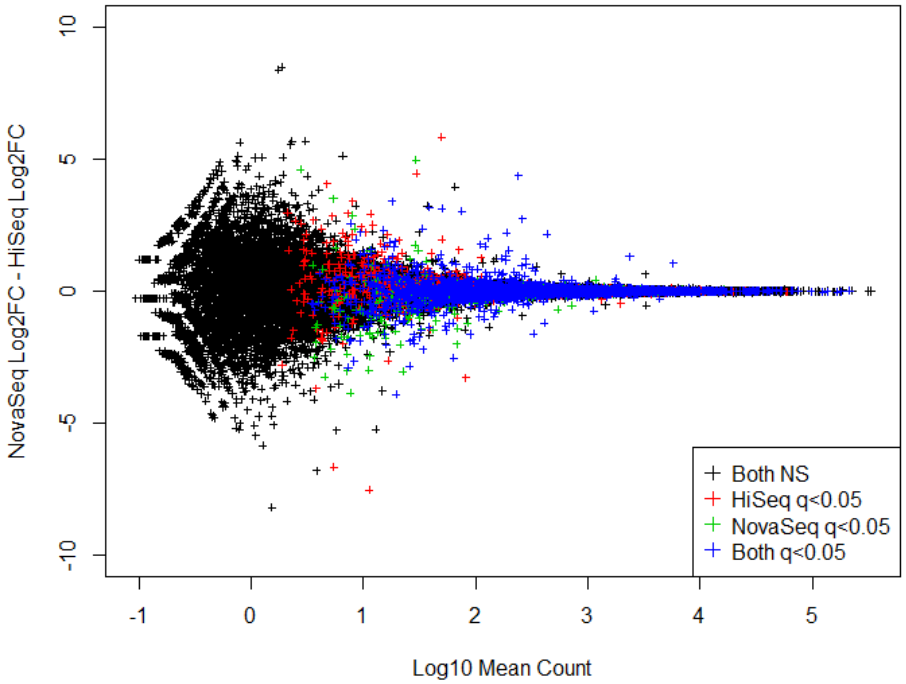
- Unsupervised hierarchical cluster analysis using the 1500 the most variable genes
- All paired samples are more similar to one another than other samples from the same instrument
- Magnitude of variation is also preserved (as before in CV plot)

Supervised Comparison



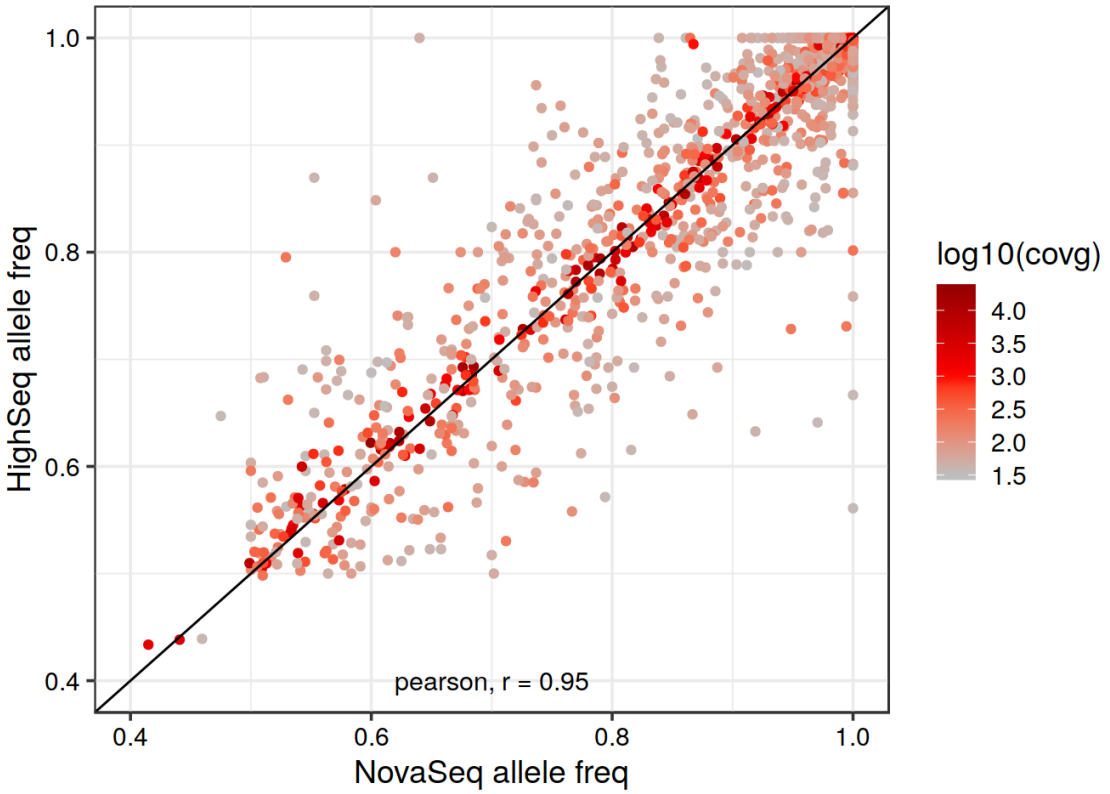


DESeq2
Claudin-low (n=4) vs Basal-like (n=4)



VAF Comparison by RNA-seq

- Exome sequencing available for one sample (Hiseq)
- All coding variants identified in DNA were quantified in RNA
- Allele frequencies by Novaseq are concordant with those of HiSeq
- High expected agreement for sequencing applications



Summary

- Systematic bias is expected when changing protocols
- Expected sources of bias – sequencing chemistry and patterned flow cells – did not broadly affect experimental results
- The magnitude of instrument bias is negligible relative observed biological variation

Acknowledgements

- Lineberger Bioinformatics

  Sara Selitsky

  David Marron

  John Ford

  Lisle Mose

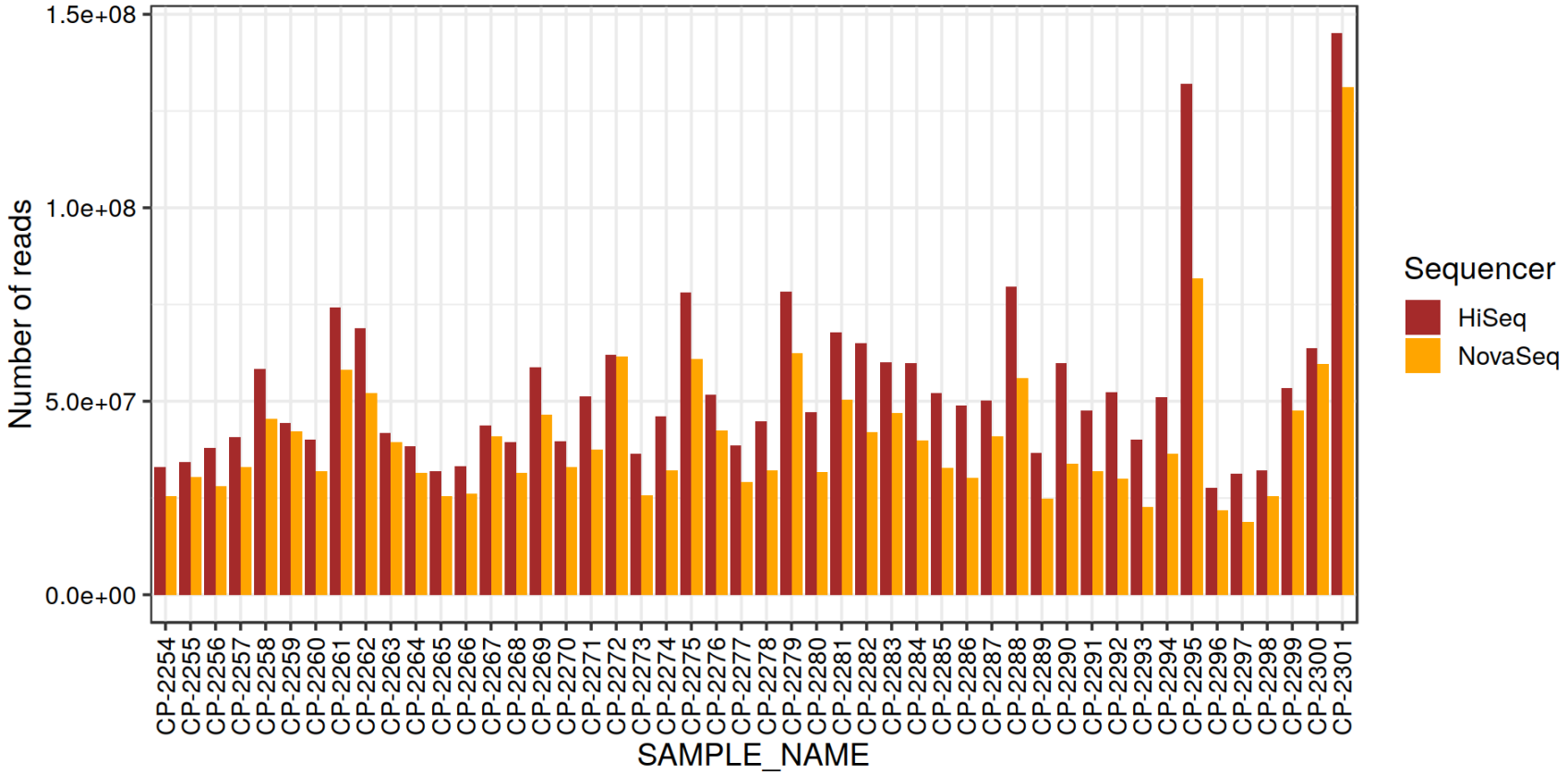
- Perou lab

  Xiaping He

  Lynn Challot

- HTSF

  Amy Perou


1.5e+08
1.0e+08
5.0e+07
0.0e+00
Number of reads
CP-2254
CP-2255
CP-2256
CP-2257
CP-2258
CP-2259
CP-2260
CP-2261
CP-2262
CP-2263
CP-2264
CP-2265
CP-2266
CP-2267
CP-2268
CP-2269
CP-2270
CP-2271
CP-2272
CP-2273
CP-2274
CP-2275
CP-2276
CP-2277
CP-2278
CP-2279
CP-2280
CP-2281
CP-2282
CP-2283
CP-2284
CP-2285
CP-2286
CP-2287
CP-2288
CP-2289
CP-2290
CP-2291
CP-2292
CP-2293
CP-2294
CP-2295
CP-2296
CP-2297
CP-2298
CP-2299
CP-2300
CP-2301
SAMPLE_NAME
Sequencer
HiSeq
NovaSeq