Within-Run Cross-Contamination in Deep Sequencing Applications on the Illumina MiSeq

22nd CROI – **Poster 592**
February 23-26, 2015
Seattle, WA, USA

Chanson J Brumme[1], Winnie Dong[1], Celia KS Chui[1], Don Kirkby[1], Richard Liang[1], Art FY Poon[1,2], P Richard Harrigan[1,2]

[1] BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada; [2] Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

Presenting Author:
Chanson Brumme
cbrumme@cfenet.ubc.ca

Background

- The Illumina MiSeq DNA sequencing system generates several gigabases of short reads per run with a relatively low error rate
- We previously described longitudinal run-to-run contamination on this platform. This carryover contamination has since been addressed by modifications to the post-run wash procedure, most notably by the addition of bleach
- Here we characterize rates and sources of systematic low level, *within-run* cross-sample contamination, an under-reported issue for this platform, and provide a potential solution

Methods

- Viral RNA or human DNA was extracted from archived plasma and whole blood samples, respectively, using a NucliSENS easyMAG
- Up to three different targets were amplified and sequenced on a single run:
  - A 327-bp fragment of HCV NS5B
  - A 266-bp fragment of HIV gp120 containing the V3 loop
  - HLA-B exons 2 (270-bp) and 3 (276-bp)
- All stages of HCV, HIV, and HLA library preparation were performed on different days by different staff
- Amplicons were dual-indexed using either barcoded PCR primers (Experiment 1), or an Illumina Nextera XT index kit (Experiment 2)
- MiSeq reads were demultiplexed with MiSeq Reporter using default settings
- Short read data were cleaned and iteratively mapped to HCV, HIV and HLA reference sequences using a custom pipeline built around bowtie2 and samtools

Experiment 1: HCV, HLA amplified with barcoded PCR primers

- In order to assess within-run cross-contamination observed in previous experiments, two libraries of disparate amplicons (HCV NS5B, human HLA-B) were sequenced at high read depth on a single MiSeq run
  - 69 amplicons (36 HCV, 33 HLA) were prepared as described above
  - 57 unique index pairs were used
  - HCV and HLA samples shared either one or two indices with samples of the opposite target (Table 1, below)
- For each sample, all recovered reads were mapped to HCV and HLA reference sequences
- "Off-target" reads were defined as HLA sequences observed in samples expected to only contain HCV (and vice versa)
- MiSeq run parameters indicated normal instrument operation and library preparation: 916 K/mm² cluster density, 88.2% reads passing filters, 84% bases >Q30

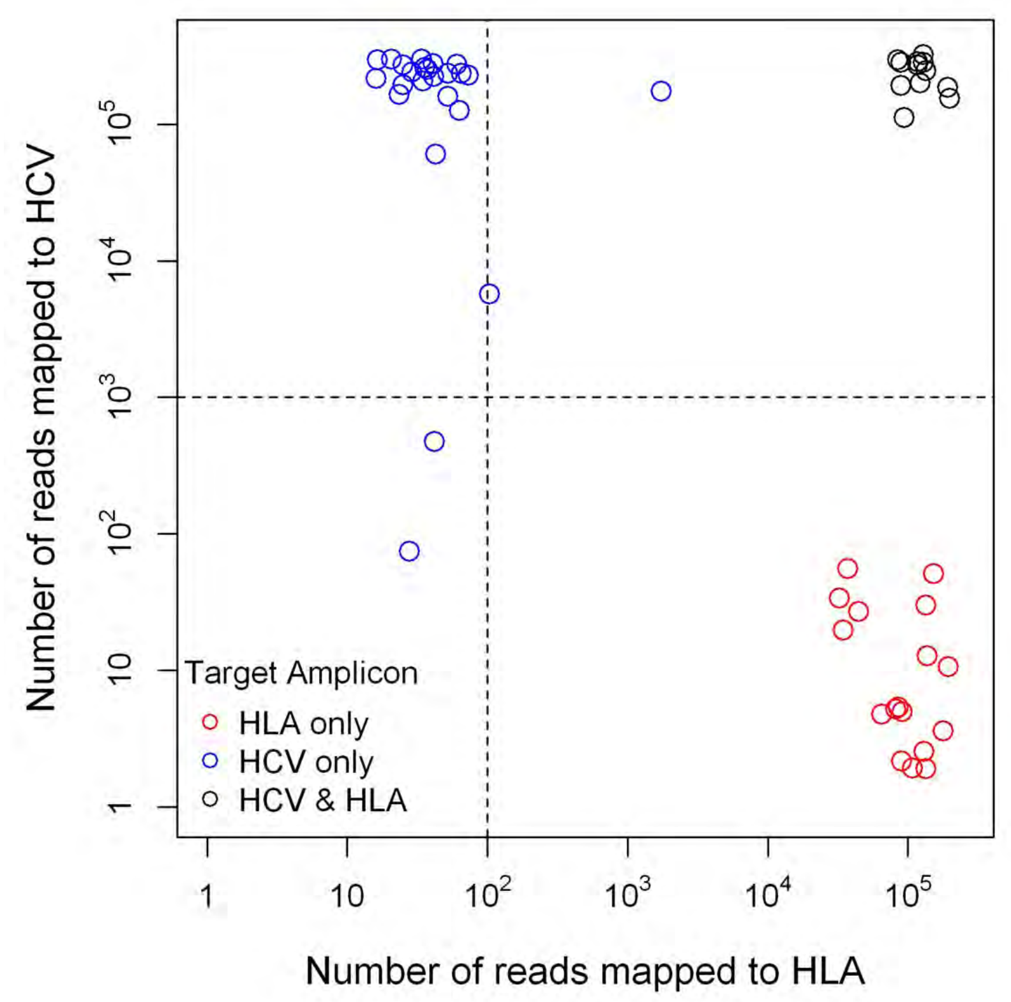
Table 1: Sample indexing strategy for Experiment 1

	N701	N702	N703	N704	N705	N706	N707	N708	N709	N710	N711	N712
N501	HLA	HLA	HLA	HLA	HLA							
N502	HCV + HLA	HCV + HLA	HCV + HLA	HCV + HLA	HCV	HCV	HCV	HCV	HCV	HCV	HCV	HCV
N503	HCV + HLA	HCV + HLA	HCV + HLA	HCV + HLA	HCV	HCV	HCV	HCV	HCV	HCV	HCV	HCV
N504	HCV + HLA	HCV + HLA	HCV + HLA	HCV + HLA	HCV	HCV	HCV	HCV	HCV	HCV	HCV	HCV
N505	HLA	HLA	HLA	HLA								
N506	HLA	HLA	HLA	HLA								
N507	HLA	HLA	HLA	HLA								
N508	HLA	HLA	HLA	HLA								

- 21 dual-index combinations (e.g. N501+N701) were used for **HLA** samples only, 24 combinations were used for **HCV** samples only and 12 combinations were used for both **HCV and HLA** samples
- All samples shared at least one index with at least one sample of the opposite target
  - For example, the **HLA** sample barcoded with N501+N701 shared the N501 index with 4 other **HLA** samples, while the N701 index was shared with 7 other **HLA** samples and 3 **HCV** samples

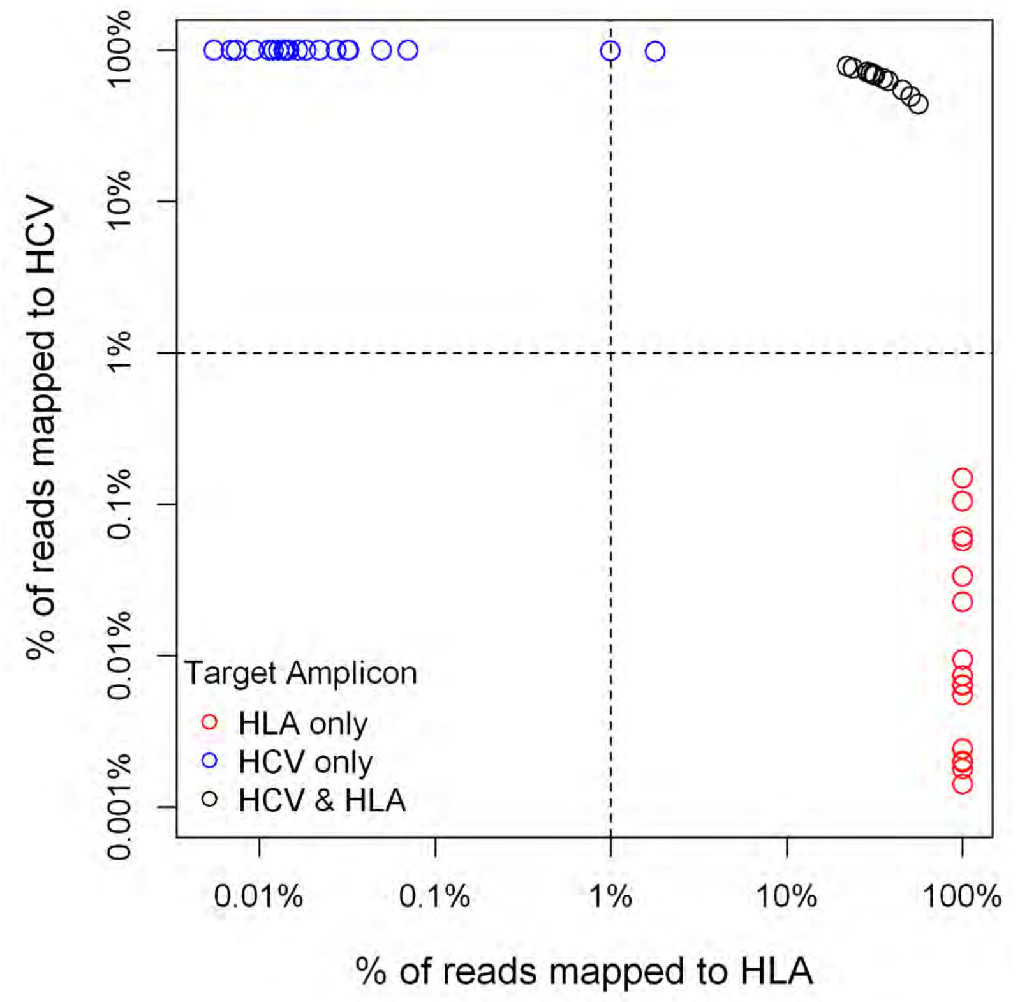
Experiment 1: Frequency of on- and off-target reads

Figure 1: Number of recovered reads that map to the HLA and HCV references



- On average, ~114,000 and ~210,000-fold coverage was obtained for HCV and HLA-B samples, respectively
- Up to 1740 HLA-B reads were observed in samples expected to contain only HCV
- Up to 56 HCV reads were observed in samples expected to contain only HLA-B
- Dashed reference lines indicate minimum coverage levels required to pass quality control

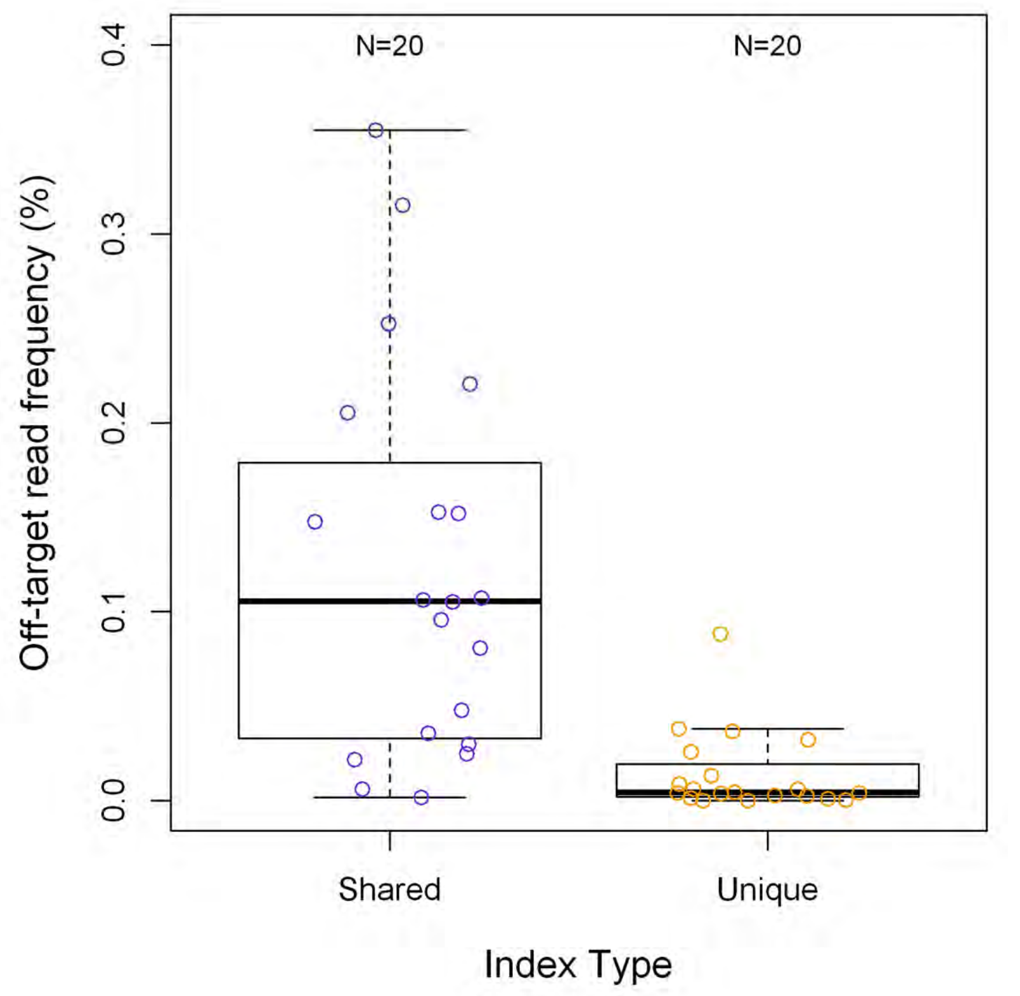
Figure 2: Percentage of recovered reads that map to the HLA and HCV references



- While only ~0.05% of all recovered reads were off-target:
  - Up to 1.8% of reads/sample were off-target
  - Off-target reads ≥1% were observed in 2 HCV samples
- Dashed reference lines indicate a typically claimed 1% limit-of-detection for low frequency variants
- Importantly, cross-contamination was also observed between samples of the same type (see Figures 4, 5)

Experiment 2: HCV, HIV, HLA indexed with Nextera XT kit

Figure 3: Frequency of off-target reads by indexing strategy



- A second experiment was performed to rule out contaminated primers, or poor primer synthesis as the source of off-target reads
- HLA-B from homozygous donors, and clonal HIV and HCV isolates were amplified separately
- Amplicons were indexed using a Nextera XT index kit (See Figures 4, 5 for indexing strategy. Not shown are 16 additional samples with unique barcodes sequenced on a separate MiSeq run)
  - Samples with shared barcodes shared a single index with another sample (7 **HCV**, 6 **HIV**, 7 **HLA** samples)
  - Samples with unique barcodes did not share any index with another sample (7 **HCV**, 7 **HIV**, 6 **HLA** samples)
- Off-target reads were more frequently observed in samples with shared vs. unique indices

Experiment 2: Identifying sources of cross-contamination

Figure 4: Frequency of contaminant reads originating from a sample with "unique" indices

	N701	N702	N703	N704	N705	N706	N707	N708	N709	N710	N711	N712
N517	0											
N502		0										
N503			120,938									
N504				0	0	0	0	0	0			
N505				0	1	0	0	0	0			
N506				0	0	0	0	0	0			
N507										0		0
N508											0	

Figure 5: Frequency of contaminant reads originating from a sample with "shared" indices

	N701	N702	N703	N704	N705	N706	N707	N708	N709	N710	N711	N712
N517	0											
N502		0										
N503			0									
N504				0	0	0	0	0	0			
N505				13	62,666	11	17	8	13			
N506				0	1	0	0	0	0			
N507										0		0
N508											0	

- Index pairs used to tag **HCV** in blue, **HIV** in green, **HLA** samples are in red
- Numbers indicate the frequency of reads matching the consensus sequence of the sample in the boxed cell
- A single read matching the consensus sequence from the unique-tagged N503-N703 sample was found in the N505-N705 sample (Figure 4)
- Multiple reads matching the consensus sequence from the shared-tagged N505-N705 sample were found in all samples indexed with N505, and one sample indexed with N705 (Figure 5)
- Similar patterns were observed for all other samples
- Screening all off-target reads against all consensus sequences indicated that the source of contamination was far more likely to be a sample that shared one index than a sample that shared none (OR=15.7, $p=10^{-11}$)

Experiment 2: Reference sequence coverage of off-target reads

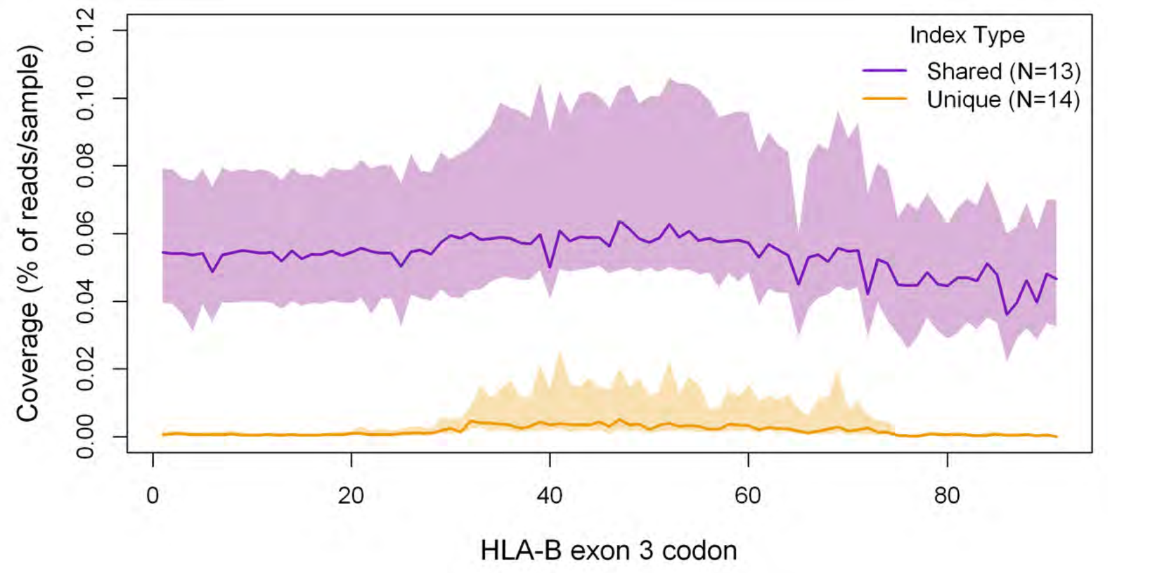


Figure 6: Off-target HLA-B read coverage

- Shown is the median (solid line) and IQR (shaded area) coverage at each position of HLA-B exon 3 from off-target HLA-B reads originating from HIV and HCV samples
- Compared to off-target reads from samples with shared indices, reads from samples with unique indices map only to the middle of exon 3. (i.e. the 5' and 3' ends are poor matches to HLA-B exon 3)
- Similar results were observed for other targets

Conclusions

- We have observed that each sample in a 96-sample run is systematically contaminated with 17 others in a fairly predictable manner. Usually the extent of cross-contamination is relatively small, but depending on the number of reads recovered per sample it can become very significant
- The source of this low-frequency cross-contamination is typically samples that share one index of the pair
- While these experiments use "off-target" reads to illustrate the issue, cross-contamination is also observed between neighboring samples of the same type
- Accurate interpretation of low-frequency variants would require knowledge of all other samples tested on the same run and bioinformatic cleanup of low-frequency contaminants
- Alternatively, this issue can be mitigated by not allowing a sample to share any indices with any other sample in the same run, a slightly more cumbersome approach