

# Common problems in Influenza bioinformatics



Ben Rambo-Martin, PhD  
Lead Bioinformatics Scientist  
US CDC – Influenza Division

# Common issues



- Failed Mira QC
  - Low-coverage / Incomplete segment coverage
    - Not enough reads in your library → Back to the lab
      - Ct  $\leq$  28?
      - Gel image resolves segment bands cleanly?
      - Proper sample number on your flow cell?
    - MIRA  $\leq$  v2.0.0 ??
      - Reads are being subsampled

# Common issues



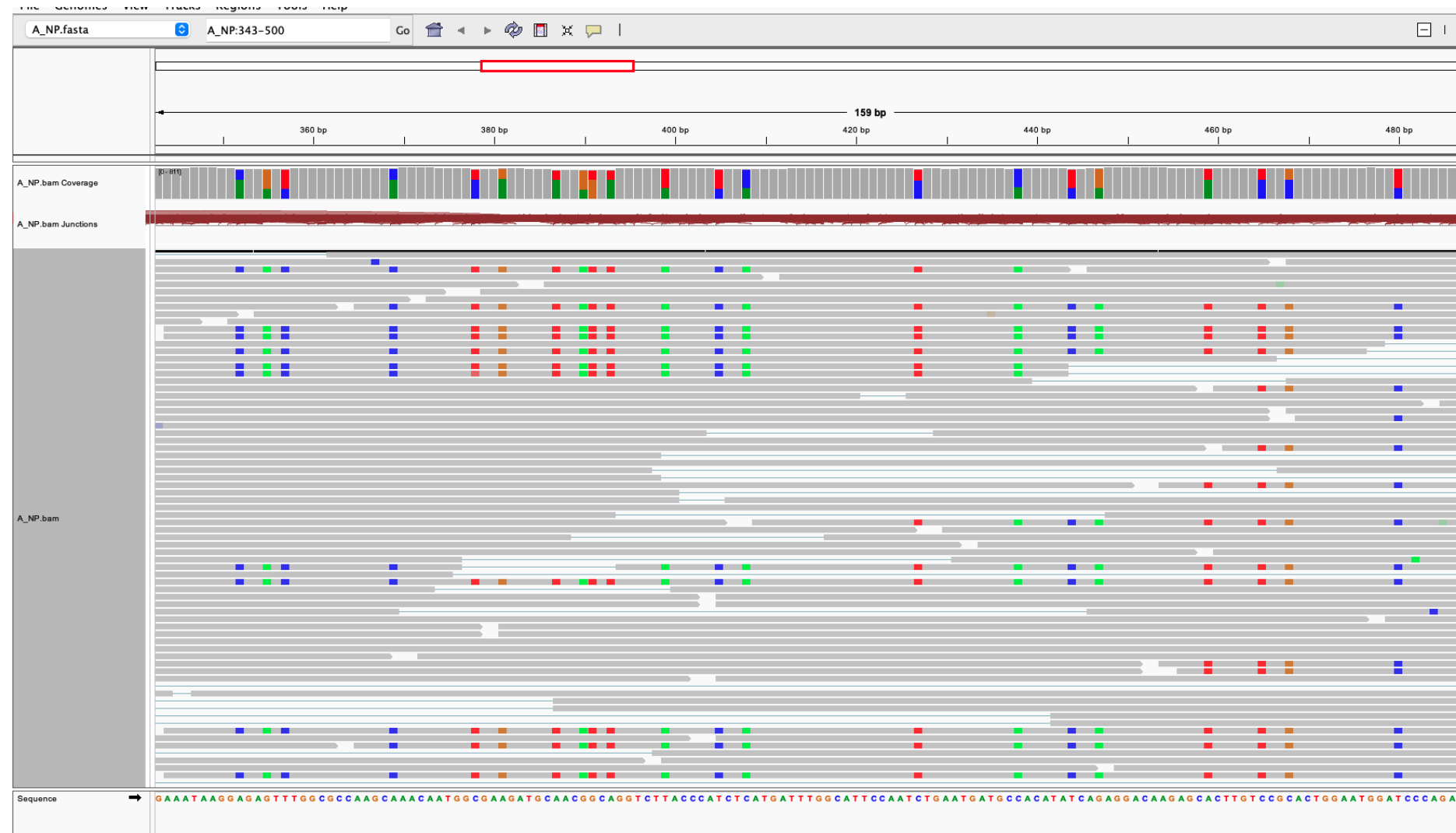
- Failed Mira QC
  - Minor variant count >10
    - In standard *clinical* samples, we have never observed >6 minor alleles ( $\geq 5\%$  frequency) with the majority having 1-3, per segment
    - Cell cultures regularly have high genetic variability which can result in high counts
    - Could be a co-infection. Unlucky person got both H1N1 and H3N2 infection *at the same time*
    - **Most likely contamination!**

sample_id	total_reads	pass_qc	reads_mapped	reference	percent_reference_coverage	median_coverage	count_minor_snv_at_or_over_5_pct	pass_fail_reason
95f48e8a	95828	95828	10388	A_HA_H1	99.82	735	2	Pass
95f48e8a	95828	95828	5818	A_MP	100	686	148	Count of minor variants at or over 5% > 10
95f48e8a	95828	95828	9528	A_NA_N1	99.79	814	3	Pass
95f48e8a	95828	95828	9146	A_NP	100	722	274	Count of minor variants at or over 5% > 10
95f48e8a	95828	95828	5704	A_NS	97.1	777	147	Count of minor variants at or over 5% > 10
95f48e8a	95828	95828	11790	A_PA	100	607	361	Count of minor variants at or over 5% > 10
95f48e8a	95828	95828	13230	A_PB1	100	687	246	Count of minor variants at or over 5% > 10
95f48e8a	95828	95828	12175	A_PB2	100	590	391	Count of minor variants at or over 5% > 10
c282a097	16170	16170	1992	A_HA_H3	99.82	158	35	Count of minor variants at or over 5% > 10
c282a097	16170	16170	1382	A_MP	100	187	12	Count of minor variants at or over 5% > 10
c282a097	16170	16170	1852	A_NA_N2	100	178	25	Count of minor variants at or over 5% > 10
c282a097	16170	16170	1720	A_NP	100	157	18	Count of minor variants at or over 5% > 10
c282a097	16170	16170	1336	A_NS	97.1	206	10	Pass
c282a097	16170	16170	2516	A_PA	100	148	22	Count of minor variants at or over 5% > 10
c282a097	16170	16170	2712	A_PB1	100	161	23	Count of minor variants at or over 5% > 10
c282a097	16170	16170	2660	A_PB2	100	160	28	Count of minor variants at or over 5% > 10

Two populations are present.

One matching reference (gray) one sharing all the colored SNVs

2+ mutations on one molecule are “phased” or “linked”



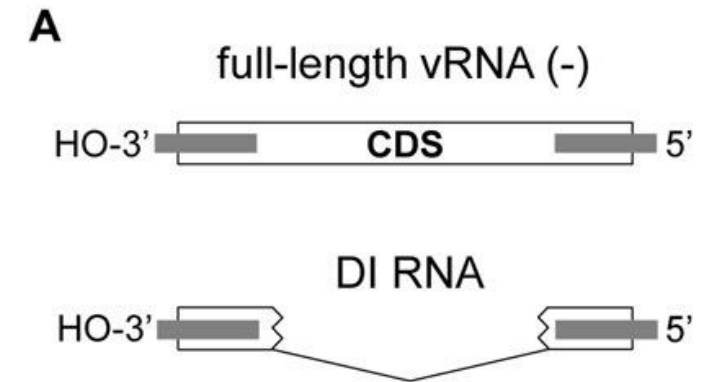
# Common issues



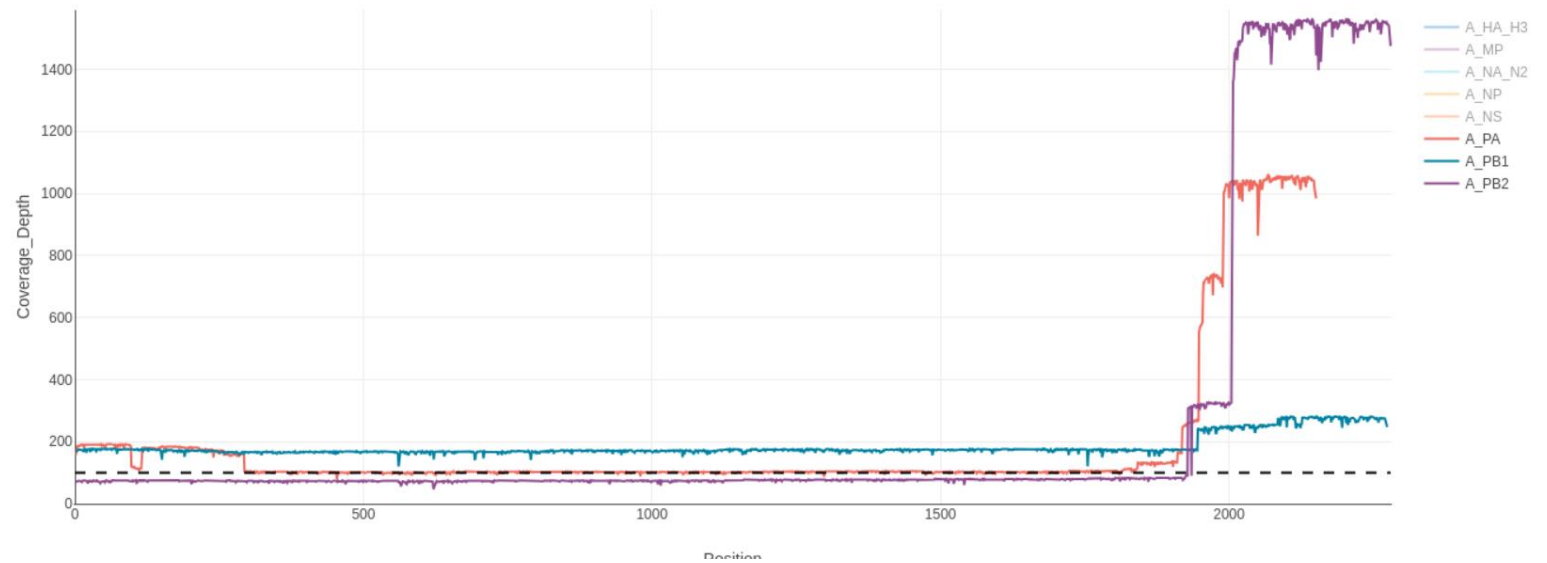
- Failed Mira QC
  - Premature stop-codon?
    - ONT homopolymer issue may require manual correction
    - DI particle induced alignment

# DI Particles

- DI Particles can interfere with NGS
- Common in polymerase segments
- Coverage shape can spike at end or “bat ears”
- Can create erroneous indel mutations at coverage dropoff points



<https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/biot.201400429>



# DI Particles



C:\Users\qgx6\Downloads\6789\_acidPB2.fasta

File Edit Select View Annotations Format Colour Calculate Web Service

630 640 650 660 670 680 690 700 710 720 730 740 750

6789|PB2/1-675 KQSRMQFSSLTVNVRGSMRILVVRGNSPVFNYNKTTKRLTILGKPEVTL  
4002251112/1-760 KQSRMQFSSLTVNVRGSGMRILVVRGNSPVFNYNKTTKRLTILGKDAGTLIEDPDESTSGVESAVLRGFLIIGKEDRRYGPALSIINELSNLAKGEKANVLIGQGDVVLVMKRRKRDSSLTDSQTATKRIRMAIN\*  
4002251111/1-760 KQSRMQFSSLTVNVRGSGMRILVVRGNSPVFNYNKTTKRLTILGKDAGTLIEDPDESTSGVESAVLRGFLIIGKEDRRYGPALSIINELSNLAKGEKANVLIGQGDVVLVMKRRKRDSSLTDSQTATKRIRMAIN\*

# Frameshifts

- Prevalent in homopolymer regions of Oxford Nanopore Sequencing
- DAIS-Ribosome is frameshift-tolerant
  - Shows as (~) mutation
- Convert back to nucleotide space and add or remove base to fix

The screenshot displays the Jalview 2.11.3.2 interface. The main window shows a sequence alignment for the file `/Users/nbx0/Downloads/nextclade.aligned (2).fasta`. The sequence is `aa/A_HA_H1/1-1700 TATTGGGGGCCATTGCCGGCTTCATTGAA~GGGGGTGGACAGGGATGGTAGATGGATGGTACG`. A red highlight is placed over the `G` in the `GAA~` region. A context menu is open, showing options like 'Selection', 'Sequence Details', 'Show annotations', 'Hide annotations', 'Add reference annotations', 'Edit', 'Output to Textbox...', 'Create Sequence Feature...', 'Create Group', and 'Edit New Group'. The 'Edit' option is selected, and a sub-menu is open showing 'Copy', 'Cut', 'Edit Sequence...', 'To Upper Case', 'To Lower Case', and 'Toggle Case'. Below the sequence, the 'Consensus' bar shows a gap corresponding to the frameshift, and the 'Occupancy' bar shows a corresponding gap. The alignment is shown with a scale from 1060 to 1110.

# Mira demands high quality results

- Together we can stop the “Garbage In” side of the data analysis mantra: “Garbage In, Garbage Out”
- Built-in thresholds are for standardizing QC
- Amended consensus gives us extra information from a single sequence.
  - A mixed site may be under active or balancing selection in the host
- Consider each sample as a whole: if HA and NA pass but all other segments fail, consider why.
- High priority samples can be useful even when QC thresholds are not met

# Mira demands high quality results

- Together we can stop the “Garbage In” side of the data analysis mantra: “Garbage In, Garbage Out”
- Built-in thresholds are for standardizing QC
- Amended consensus gives us extra information from a single sequence.
  - A mixed site may be under active or balancing selection in the host
- Consider each sample as a whole: if HA and NA pass but all other segments fail, consider why.
- High priority samples can be useful even when QC thresholds are not met

# Mira demands high quality results

- Together we can stop the “Garbage In” side of the data analysis mantra: “Garbage In, Garbage Out”
- Built-in thresholds are for standardizing QC
- Amended consensus gives us extra information from a single sequence.
  - A mixed site may be under active or balancing selection in the host
- Consider each sample as a whole: if HA and NA pass but all other segments fail, consider why.
- High priority samples can be useful even when QC thresholds are not met

# Mira demands high quality results

- Together we can stop the “Garbage In” side of the data analysis mantra: “Garbage In, Garbage Out”
- Built-in thresholds are for standardizing QC
- Amended consensus gives us extra information from a single sequence.
  - A mixed site may be under active or balancing selection in the host
- Consider each sample as a whole: if HA and NA pass but all other segments fail, consider why.
- High priority samples can be useful even when QC thresholds are not met

# Mira demands high quality results

- Together we can stop the “Garbage In” side of the data analysis mantra: “Garbage In, Garbage Out”
- Built-in thresholds are for standardizing QC
- Amended consensus gives us extra information from a single sequence.
  - A mixed site may be under active or balancing selection in the host
- Consider each sample as a whole: if HA and NA pass but all other segments fail, consider why.
- High priority samples can be useful even when QC thresholds are not met