The Turkana Genome Project
WGS Data Analysis

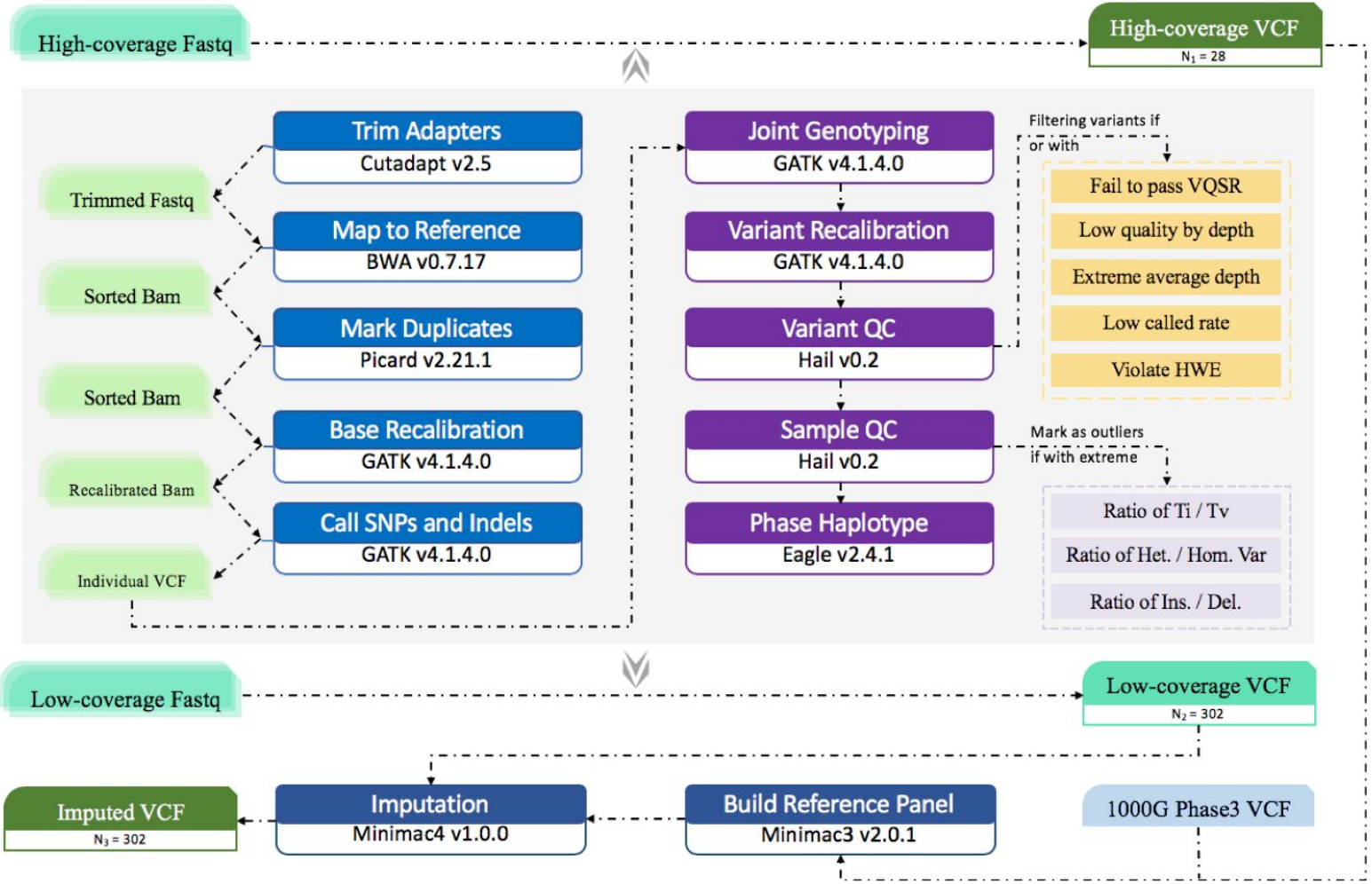
Yushi Tang (*Ayroles Lab*)

Dec 6, 2019

Part I: Mapping WGS to Human Genome

- From FASTQ to VCFs
- How to process high-coverage WGS raw data?
  - References: https://software.broadinstitute.org/gatk/best-practices/workflow?id=11165
  - Source code: https://github.com/YushiFT/turkana-genome
- How to process low-coverage WGS data ?
  - How to improve imputation accuracy for low-coverage samples?
  - What is the most idea imputation reference panel (IRP) for Turkana project?
  - Literature review
    - References: https://www.dropbox.com/home/Turkana_genomics/imputation_papers

Pipeline


High-coverage Fastq
High-coverage VCF
N1 = 28
Trim Adapters
Cutadapt v2.5
Trimmed Fastq
Map to Reference
BWA v0.7.17
Sorted Bam
Mark Duplicates
Picard v2.21.1
Sorted Bam
Base Recalibration
GATK v4.1.4.0
Recalibrated Bam
Call SNPs and Indels
GATK v4.1.4.0
Individual VCF
Joint Genotyping
GATK v4.1.4.0
Variant Recalibration
GATK v4.1.4.0
Variant QC
Hail v0.2
Sample QC
Hail v0.2
Phase Haplotype
Eagle v2.4.1
Filtering variants if
or with
Fail to pass VQSR
Low quality by depth
Extreme average depth
Low called rate
Violate HWE
Mark as outliers
if with extreme
Ratio of Ti / Tv
Ratio of Het. / Hom. Var
Ratio of Ins. / Del.
Low-coverage Fastq
Low-coverage VCF
N2 = 302
Imputed VCF
N3 = 302
Imputation
Minimac4 v1.0.0
Build Reference Panel
Minimac3 v2.0.1
1000G Phase3 VCF

Turkana Genome Program (28 High-Coverage Samples)

Cohort VCF --> VQSR + Phasing: 24,107,283 Records, 18,341,801 SNPs, 5,176,650 Indels

Chrom.	1	2	3	4	5	6	7	8	9	10	11
SNPs	1408838	1504002	1281658	1288066	1133242	1114258	1077580	997129	851694	918,323	891171
Indels	420211	424000	351615	345612	315013	319191	306682	260491	222703	258,109	235470
All	1873583	1974566	1670070	1675010	1481756	1468888	1421073	1287012	1098116	1,208,344	1152057
Chrom.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
SNPs	846328	610569	574463	558251	608439	540141	514636	429419	494126	340038	359430
Indels	254141	176911	168197	161057	170759	183099	138212	155869	129728	86137	93443
All	1127836	807025	760261	736583	800130	745601	668479	603299	641232	439607	466755

Turkana Genome Program (28 High-Coverage Samples)

Cohort VCF --> VQSR + Phasing: Ti / Tv = 1.93 (12,093,566 transitions and 6,248,235 transversions)

Chrom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Transition	942702	989415	839199	845677	746933	746469	710058	647223	552367	606,783	583047
Transver.	466136	514587	442459	442389	386309	367789	367522	349906	299327	311,540	308124
Ti / Tv	2.02	1.92	1.90	1.91	1.93	2.03	1.93	1.85	1.85	1.95	1.89
Chrom.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Transition	560921	408899	385280	369789	390413	363366	338368	290007	321116	219381	236153
Transver.	285407	201670	189183	188462	218026	176775	176268	139412	173010	120657	123277
Ti / Tv	1.97	2.03	2.04	1.96	1.79	2.06	1.92	2.08	1.86	1.82	1.92

Turkana Genome Program (28 High-Coverage Samples)

Cohort VCF --> VQSR + Phasing + Imputation: 47,109,465 Records, 43,835,856 SNPs, 3,233,842 Indels

Chrom.	1	2	3	4	5	6	7	8	9	10	11
SNPs	3478709	3777037	3120273	3098395	2816013	2740366	2565019	2484282	1925625	2174629	2175376
Indels	256605	277300	232822	237167	213854	211538	185918	165158	135621	157600	155857
All	3738240	4057613	3355939	3338265	3032422	2954410	2753497	2651561	2063096	2334090	2333242
Chrom.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
SNPs	2079499	1537173	1425799	1305267	1455657	1248577	1228485	1008193	978041	606264	607177
Indels	161228	123054	108524	97688	92390	96116	89899	75354	68716	47006	44427
All	2242720	1661700	1535592	1404164	1549316	1345835	1319629	1084535	1047613	653791	652195

Turkana Genome Program (28 High-Coverage Samples)

Cohort VCF --> VQSR + Phasing + Imputation: Ti / Tv = 2.16 (29,980,399 Ti. and 13,855,309 Tv.)

Chrom.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Transition	2406115	2570900	2118877	2091514	1919000	1877141	1751090	1656806	1296178	1498448	1486886
Transver.	1072584	1206124	1001386	1006874	897003	863219	813919	827465	629442	676170	688483
Ti / Tv	2.24	2.13	2.12	2.08	2.14	2.17	2.15	2.00	2.06	2.22	2.16
Chrom.	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Transition	1434167	1050929	980414	892215	970415	885246	845382	710726	687261	418550	432139
Transver.	645326	486241	445378	413047	485238	363323	383097	297462	290777	187713	175038
Ti / Tv	2.22	2.16	2.20	2.16	2.00	2.44	2.21	2.39	2.36	2.23	2.47

Part II: Processing Unmapped Reads

- Acquire unmapped reads:
  - Raw FASTQ -> Trim Adpters -> BWA -> Unmapped Reads
  - Mapping Unmapped Reads: NCBI Magic-BLAST (v1.5.0, https://ncbi.github.io/magicblast/)
- Mapping unmapped reads to NCBI reference genome:
  - 1. Plasmodium falciparum (malaria parasite P. falciparum)
    - One of the causal agents of human malaria
    - https://www.ncbi.nlm.nih.gov/genome/?term=plasmodium+3d7
  - 2. Plasmodium vivax (malaria parasite P. vivax)
    - Second most important causal agent of human malaria
    - https://www.ncbi.nlm.nih.gov/genome/35

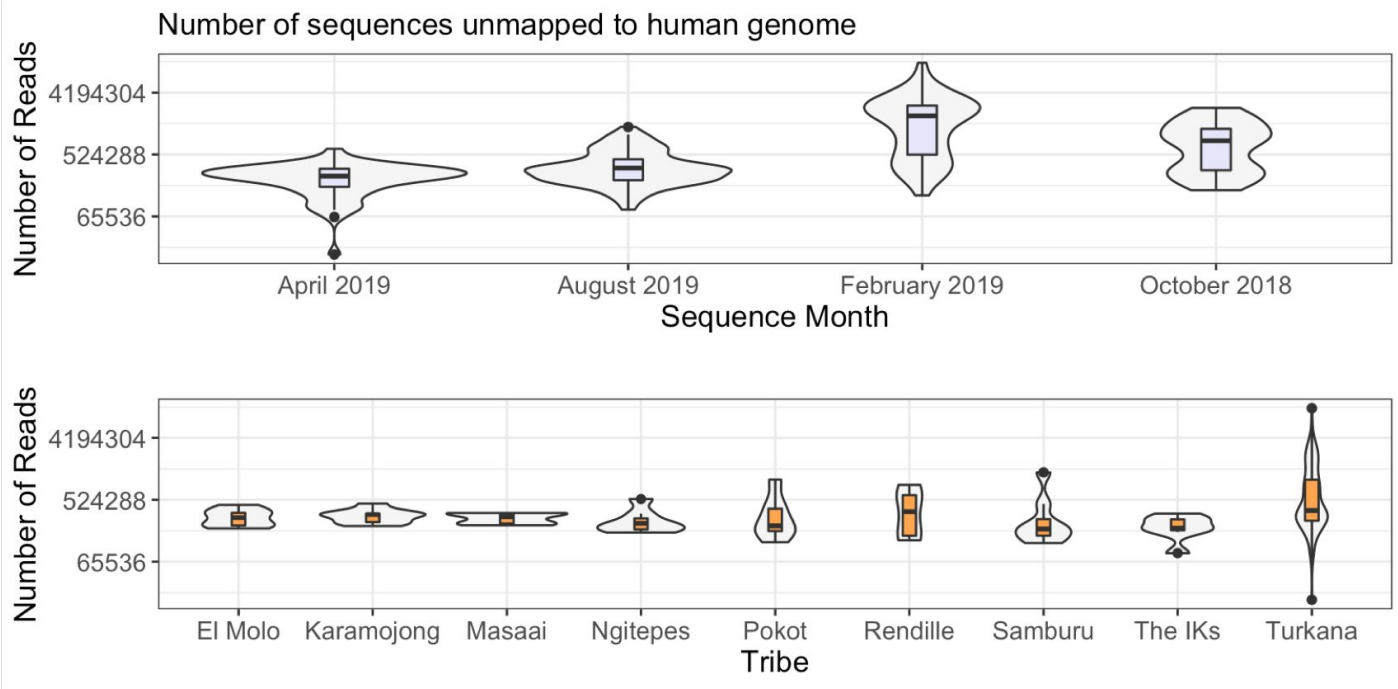
Part II: Processing Unmapped Reads

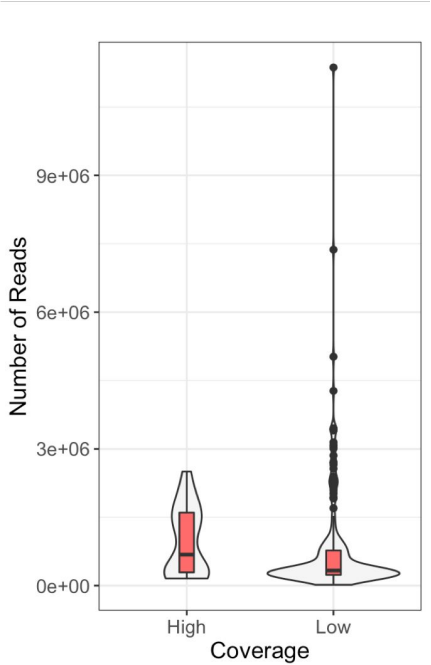
- Mapping unmapped reads to NCBI reference genome (*con't*):
  - 1. Plasmodium falciparum (malaria parasite P. falciparum)
  - 2. Plasmodium vivax (malaria parasite P. vivax)
  - 3. Mycobacterium tuberculosis
    - Causative agent of tuberculosis
    - https://www.ncbi.nlm.nih.gov/genome/?term=mycobacterium+tuberculosis+H37Rv
  - 4. Pasteurella multocida
    - Pathogen in human, swine and poultry, an opportunistic pathogen that causes cholera
    - https://www.ncbi.nlm.nih.gov/genome/912
  - 5. Trichuris trichiura (human whipworm): https://www.ncbi.nlm.nih.gov/genome/?term=Trichuris+trichiura

Part II: Processing Unmapped Reads

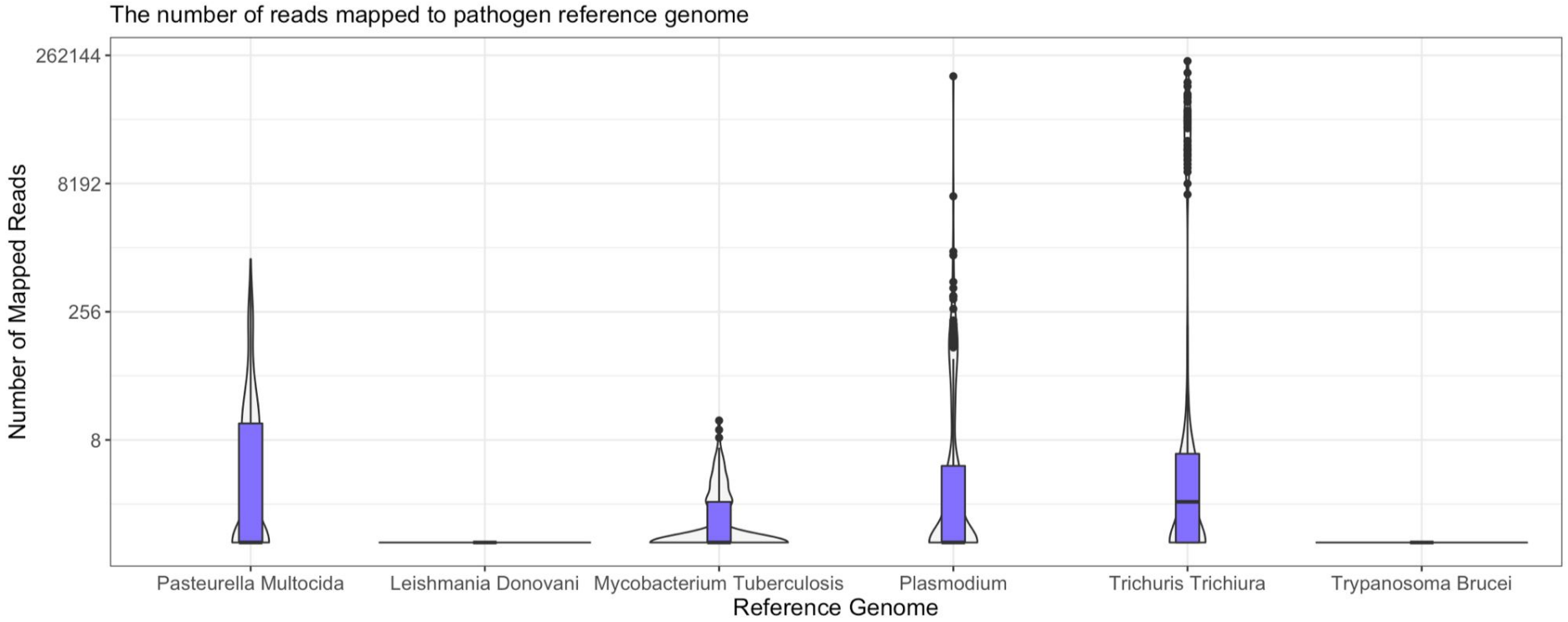
- Mapping unmapped reads to NCBI reference genome (*con't*):
  - 1. Plasmodium falciparum (malaria parasite P. falciparum)
  - 2. Plasmodium vivax (malaria parasite P. vivax)
  - 3. Mycobacterium tuberculosis
  - 4. Pasteurella multocida
  - 5. Trichuris trichiura (human whipworm)
  - 6. Leishmania donovani
    - Agent for human leishmaniasis
    - https://www.ncbi.nlm.nih.gov/genome/3516
  - 7. Trypanosoma brucei
    - Unicellular flagellated protozoan that is the cause of sleeping sickness
    - https://www.ncbi.nlm.nih.gov/genome/24

The number of sequences unmapped to human genome varies mostly across different sequencing month

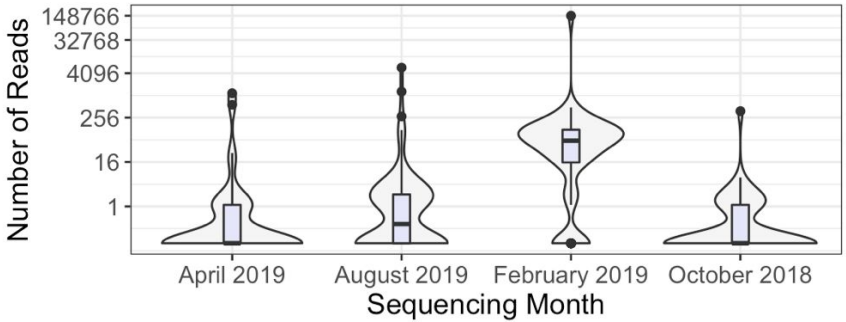


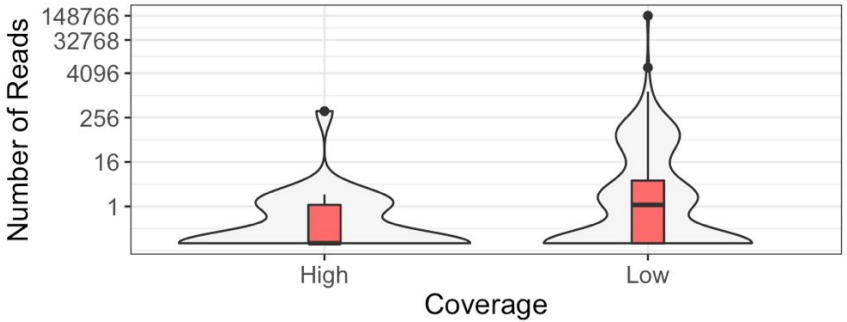


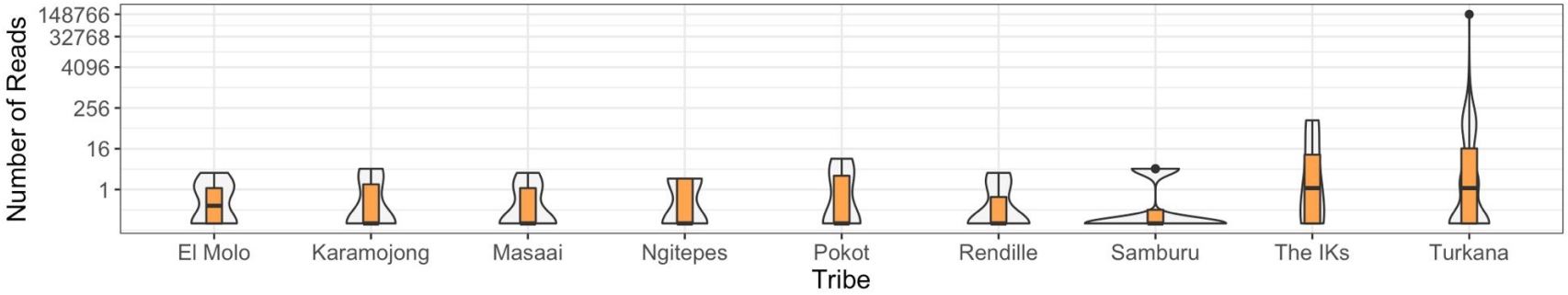
Partial of unmapped human sequences are aligned to four pathogen reference genomes (N = 329)



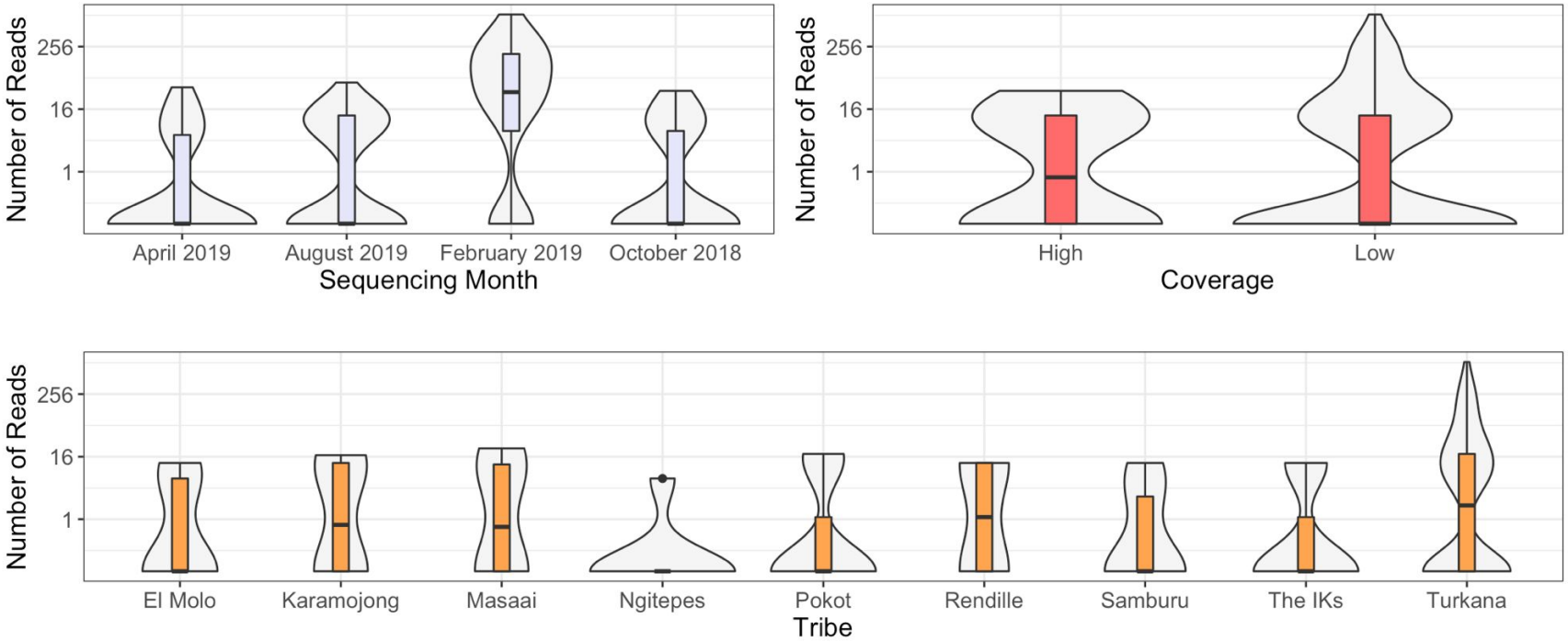
Plasmodium (malaria parasite): 133 samples were detected to carry the plasmodium pathogen (N = 329)



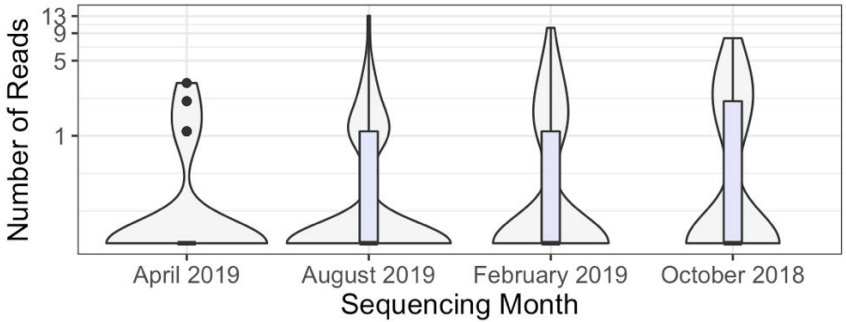


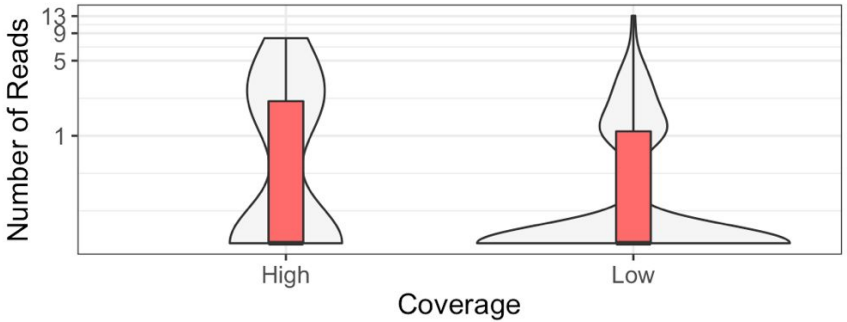


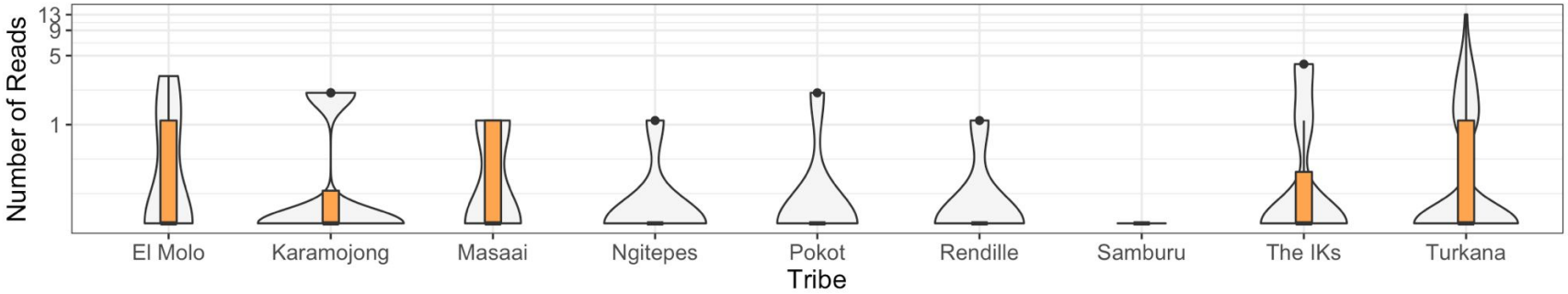
Pasteurella multocida: 128 samples were identified to carry the pasteurella multocida pathogen (N = 329)



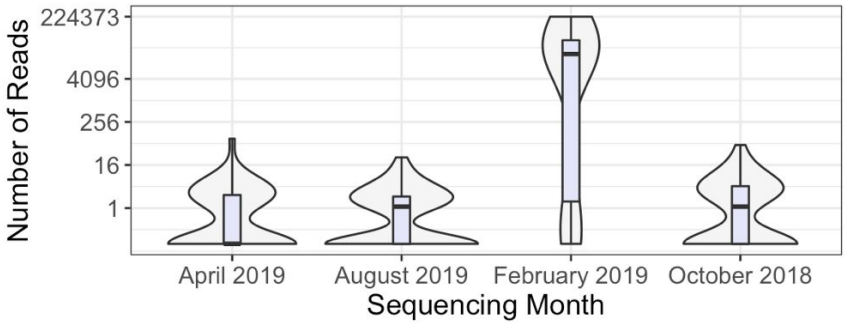
Mycobacterium tuberculosis: 76 samples were identified to carry the mycobacterium tuberculosis pathogen (N = 329)

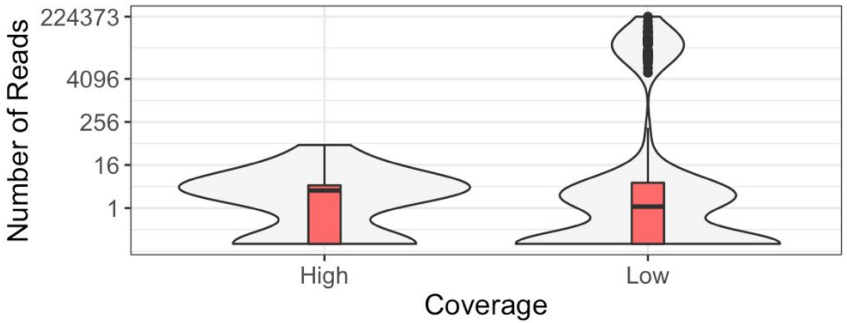


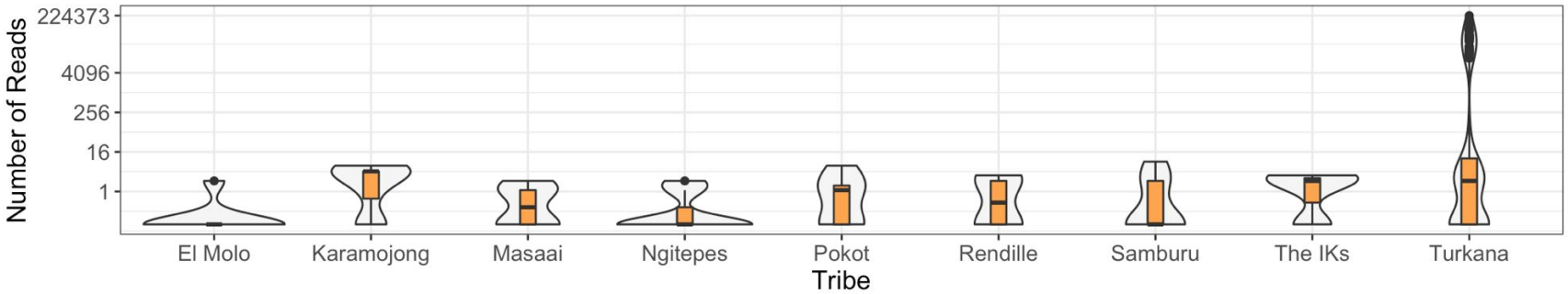




Trichuris trichiura: 153 samples were identified to carry the trichuris trichiura pathogen (N = 329)







Acknowledgment

- Ayroles Lab at Princeton University
  - Prof. Julien F. Ayroles
  - Dr. Amanda J. Lea
- Lewis-Sigler Institute of Integrative Genomics, Princeton University
  - Lance R. Parsons
- *Thanks for your attention!*