Long-read sequencing to interrogate strain-level variation among adherent-invasive *Escherichia coli* isolated from human intestinal tissue

Jeremy Wang[1¶], Rachel Bleich[2¶], Sandra Zarmer[3], Janelle Arthur[2,4,5*]

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2]Department of Microbiology & Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3]Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4]Center for Gastrointestinal Biology & Disease, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[5]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

¶ These authors contributed equally to this work.

* Corresponding author

Email: janelle_arthur@med.unc.edu

Abstract

Adherent-invasive *Escherichia coli* (AIEC) are a pathovar linked to inflammatory bowel diseases (IBD), especially Crohn's disease, and colorectal cancer. AIEC have no known molecular or genomic markers, but instead are defined by *in vitro* functional attributes. Futhermore, it is unknown if strains classified as AIEC truly colonize intestinal tissues better than non-AIEC strains. To evaluate strain-level variation among tissue-associated *E. coli*, we must develop a sequencing approach capable of long reads and with the ability to exclude mammalian DNA. We also must evaluate genomic variation among strains that have demonstrated ability to colonize intestinal tissues. Here we have assembled complete genomes using ultra-long-read nanopore sequencing for a model AIEC strain, NC101, and seven strains isolated from the intestinal mucosa of Crohn's disease and non-Crohn's tissues. We show these strains can colonize the intestinal tissue in a Crohn's disease mouse model and induce varying levels of inflammatory cytokines from cultured macrophages. We demonstrate these strains can be quantified and distinguished in the presence of 99.5% mammalian DNA and from within a fecal population. Analysis of global genomic structure and specific sequence variation within the ribosomal RNA operon provides a framework for efficiently tracking strain-level variation of closely-related *E. coli* and likely other commensal/pathogenic bacteria impacting intestinal inflammation in mice and IBD patients.

Introduction

Crohn's disease (CD) is a type of Inflammatory bowel disease (IBD), a chronic inflammatory condition that is the result of an inappropriate immune response towards elements of the intestinal microbiota [1]. Due to chronic inflammation and exposure to pro-carcinogenic microbes, CD patients are at a greater risk of developing colorectal cancer (CRC) [2-4]. Dysbiosis of gut microbial communities in CD patients has been established, including an increase of adherent/invasive *E. coli* (AIEC) adherent to the intestinal tissues, termed the mucosal niche [5-9] . AIEC induce inflammation in experimental mouse models of IBD, including the well-established inflammation susceptibe *Il10*$^{-/-}$ mouse model [3, 10-12]. These non-toxigenic AIEC strains are defined through the *in vitro* ability to adhere to and invade epithelial cells and survive in macrophages [5, 7, 13]. However, no genomic definition distinguishes AIEC, limiting our ability to detect these pro-inflammatory strains from among patient microbiota.

Escherichia coli make up a diverse species that is known to variously have beneficial, commensal, or pathogenic impact in the mammalian gut. However, most commonly used methods for characterizing the composition of gut microbiota from fecal or tissue samples, notably sequencing of selected variable regions of the 16S rRNA operon, do not differentiate among these relatively closely-related but functionally divergent types. Emerging single-molecule (third-generation/long-read) sequencing technologies including Pacific Biosciences ("PacBio") and Oxford Nanopore Technologies ("nanopore") produce reads 10s to 100s of kilobases in length, often without prior amplification. Recent work [14-17] has demonstrated the utility of these longer reads for sequencing the entire 16S gene or entire rRNA operon, with

corresponding increase in taxonomic resolution by capturing more variable sequence, including all nine variable regions of 16S, the internal transcribed spacers (ITS), and 23S. To support the extension of these approaches to characterize strain-level variation of tissue-associated (mucosa-associated) microbiota contributing to IBD and in models of experimental colitis in mice, with a particular focus on mechanistic studies of AIEC, we produced complete genome assemblies for eight AIEC and non-AIEC *E. coli*, describe the genomic variation among these strains, particularly within the rRNA operon, and demonstrate the accurate identification of these strains in mixed *in vitro* and *in vivo* microbiota.

Methods

DNA extraction and nanopore sequencing

We extracted ultra-high-molecular-weight genomic DNA from liquid E. coli cultures using a modified phenol:chloroform protocol [18]. Briefly, 1-1.5ml of stationary-phase liquid culture was pelleted and resuspended in TE/NaCl/Triton buffer. Cells were lysed by addition of SDS to a final concentration of 2% and Proteinase K to a final concentration of 300mg/ml and incubated at 55C for 30 minutes. DNA was purified twice by addition of 1x volume phenol:chloroform and phase separation. DNA was precipitated by addition of 0.1x volume sodium acetate and 2x volume isopropanol. Ultra-high-molecular weight DNA was “hooked” out with a melted glass capillary and washed in 70% ethanol, dried, and resuspended in nuclease-free water (NFW) by incubation overnight at 4C. Multiplexed sequencing libraries were prepared for nanopore sequencing using Oxford Nanopore’s Ligation Sequencing Kit (LSK109) and Barcoding Expansion (NBD104) per the manufacturer’s recommended protocol with the following modifications [19]. Instead of SPRI bead cleanup after each ligation reaction, we added 1 volume

of "clumping buffer" (9% PEG 8000, 1M NaCl, 10mM Tris), incubated at room temperature for 30 mins, pelleted by centrifugation at 13K rpm, washed twice with 70% ethanol, then allowed to dry and resuspended in NFW by incubating at 4C overnight. This "bead-free" cleanup both avoids shearing and loss of high-molecular-weight DNA and selects for larger fragments in the PEG precipitations. Multiplexed samples were sequenced on R9.4.1 flow cells on either MinION or GridION with real-time basecalling using Guppy v3.2.2 in high-accuracy mode. The distribution of read lengths is shown in Figure 1.

Fig 1. Read length distribution across samples using high-molecular weight extraction and bead-free library prep.

Genome assembly

Basecalled reads for each sample were assembled using Miniasm v0.2 [20], followed by four rounds of polishing with Racon v1.3.1 [21], and Medaka v0.10.0 [22]. In all cases, the polished assembly consisted of one contig representing the full-length genome. Assembled genomes were normalized to start at the origin of replication and re-polished across the previous breakpoint. Annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline [PGAP release 2020-02-06.build4373; 23]. Predicted genes were subsequently assigned putative function by aligning to the RefSeq non-redundant protein database with Diamond v0.9.22 [24] allowing for frame-shift sensitive alignment (diamond blastp --more-sensitive --frameshift 15).

Serotype, MLST, virulence, and antibiotic resistance genes

Polished assemblies were aligned against EcOH database [25] for serotyping, MLST and VirulenceFinder databases [26], and NCBI antimicrobial resistance database using Minimap2 [27] with parameter '-cx asm5' to allow for indel-sensitive alignment appropriate for nanopore-assembled genomes. For serotype and MLST, the single best-mapping type gene (or genes, in a tie) is given as the predicted type.

rRNA operon analysis

All copies of the rRNA operon in each assembly were identified by aligning to the *E. coli* K12 reference rRNA using Minimap2 [27]. Variants relative to the K12 reference were identified for each assembled rRNA copy across all assemblies based on the alignments. Total polymorphism distance was computed between every pair of alignments across each region of the rRNA operon (16S, ITS, 23S).

Bacterial Growth Conditions

Bacteria were grown in Luria broth (LB) at 37°C and 250 rpm unless otherwise indicated.

Animal care

Germ-free mice were reared in the National Gnotobiotic Rodent Resource Center at UNC Chapel Hill. All animal experiments and procedures were approved by UNC's Institutional Animal Care and Use Committee (IACUC).

Murine stool samples

Three (2m/1f) interleukin-10-deficient (*Il10*$^{-/-}$) mice (129Sv/Ev background) were reared germ-free to adulthood (8-10) weeks and colonzied by oral gavage with an even mixture of a total of 10^7 CFU clinical *E. coli* strains isolated from the intestinal tissue of Crohn's disease and non-Crohn's disease patients [7]. All strains have been classified as AIEC and non-AIEC using standard *in vitro* assays to evaluate adhesion/invasion to Caco2 epithelial cells and uptake/survival in J774 macrophages [5,7]. To validate our results by tracking individual strains with PCR, each strain was marked with an antibiotic resistance cassette and molecular barcode inserted into a neutral chromosomal region [28-29]. These strains are named by their laboratory designation and barcode: JA0018/A1, JA0019/A3, JA0022/B6, JA0036/C5, JA0044/D2, JA0048/D5, and JA0091/C2. After colonization, mice were maintained in specific pathogen free (SPF) housing, where they will acquire a simplified mouse microbiota. After 2 weeks of colonization, we gave kanamycin water *ad libetum* for 2 weeks to suppress this microbiota and ensure that all strains could persist to some extent. Mice were sacrificed 6 weeks later (by CO_2 asphyxiation), for a total of 10 weeks colonization. A stool sample was removed from the lumen of the distal colon for our analysis. DNA was extracted and purified as described in [3,30].

Lab Strain	Strain	AIEC	Barcode
JA0018	39ES-1	non-AIEC	A1
JA0019	532-9	non-AIEC	A3
JA0022	LF82	AIEC	B6
JA0036	568-3	AIEC	C5
JA0044	37RT-2	non-AIEC	D2
JA0048	42ET-1	AIEC	D5
JA0091	HM670	non-AIEC	C2
NC101	Murine[3,10]	AIEC	-

Table 1: Strains used in study.

In vivo colonization and persistence studies

To demonstrate that the seven human-derived strains and murine *E. coli* NC101 could colonize germ-free *Il10*$^{-/-}$ mice throughout the gastrointestinal tract, we singly housed 1 male and 1 female mouse per *E. coli* strain, for a total of 16 cages. Mice were gavaged with 10^8 CFU of a single strain and moved to SPF housing. After 1 week, each mouse received a fecal transplant (prepared anaerobically from a pool of 7 C57BL/6 WT mice) via gavage to provide niche competition. Stool samples were collected almost daily and CFUs were quantified on kanamycin plates to monitor fecal *E. coli* colonization. Mice were sacrificed by CO_2 asphyxiation after 5 weeks, and the following tissues were harvested and CFUs quantified on kanamycin plates to measure viable colonizing bacteria: ileal tissue, cecal content, colon content, colon tissue and colon mucus layer. *E. coli* CFUs were also quantified from stomach content, duodenal content, and jejunal content, but only 1-3 of each tissue had detectable bacterial growth at less than 10^4 CFUs/10 mg.

To demonstrate that consistent colonizing strain, JA0048/D5, and inconsistent colonizing strain, JA0091/C2, colonized germ-free *Il10*$^{-/-}$ mice in a similar manner across a larger cohort, we housed 7 (3M/4F) or 8 (4M/4F) mice, respectively, in two cages for each cohort. Mice were gavaged with 10^8 CFU of a single strain and moved to SPF housing. After 1 week, each mouse received a fecal transplant (prepared anaerobically from a pool of 7 C57BL/6 WT mice) via gavage to provide niche competition. After 3 weeks of colonization (2 weeks post-FMT), we gave kanamycin water *ad libetum* for 2 weeks to suppress this microbiota and ensure that all strains could persist to some extent. Mice were sacrificed 6 weeks later (by CO_2 asphyxiation), for a total of 10 weeks colonization. The following tissues were harvested and CFUs quantified

on kanamycin plates to measure viable colonizing bacteria: colon content, colon tissue and colon mucus layer.

In vitro co-culture assays to measure pro-inflammatory cytokine production

To assess the inflammatory capacity of the clinical strains, J774 macrophages grown in DMEM complete media (supplemented with 10% heat-inactivated FBS and 100 U/ml penicillin-100 mg/ml streptomycin) were transferred to 12-well plates at $5x10^5$ cells/well. Cells were grown overnight at 5% CO_2 at 37°C, and washed twice with PBS before adding fresh culture media without antibiotics. Cells were infected at an MOI of 1 for 4 hours before removing the bacteria, washing with PBS, and adding fresh media with 100 µg/mL gentamycin. Cells were grown another 20 hours at which point the supernatants were collected for ELISA and the cells were collected for qPCR analysis.

ELISA

Supernatants from stimulated J774 macrophages (above) were analyzed for cytokine IL12p40 production through ELISA following manufacturer's protocol (BD Bioscience Opt EIA, catalog #555165). Assayed in 4 independent experiments with 2-3 technical replicate wells.

Quantification of cytokine expression by qPCR

Stimulated J774 macrophages were washed with PBS and transferred into 1 mL Trizol (Invitrogen) and RNA was extracted following the manufacturer's protocol. Isolated RNA was subjected to DnaseI treatment (Invitrogen) prior to cDNA synthesis. cDNA synthesis was completed using qScript cDNA SuperMix (Quantabio). qPCR amplification was performed in

triplicate with SYBR green qPCR chemistry (Bioline) using primers for *Tnfa* (F-5'-ACCTCACACTCAGATCATCTTCTC-3', R-5'-TGAGATCCATGCCGTTGG-3'), *Il1b* (F-5'-ACAGAATATCAACCAACAAGTGATATTCTC-3', R-5'-GATTCTTTCCTTTGAGGCCCA-3'), *Il12B(p40)* (F-5'-CGCAAGAAAGAAAAGATGAAGGAG-3', R-5'-TTGCATTGGACTTCGGTAGATG-3'), and *Gapdh* (F-5'-GGTGAAGGTCGGAGTCAACGGA-3', R-5'-GAGGGATCTCGCTCCTGGAAGA-3') on a QuantStudio 6 Real-Time PCR System. C_t values were normalized to *Gapdh* to generate ΔC_t values, and fold changes were calculated by $\Delta\Delta C_t$ to the ΔC_t of unstimulated controls. Assayed 3 independent experiments in triplicate.

Quantification of strains by qPCR

Stool DNA (6 ng each) from three mice was subjected to qPCR with primer pairs targeting each barcode to quantify relative amounts of each barcoded strain from within a complex population. Amplification was performed in duplicate using SYBR green qPCR chemistry (Bioline) using a universal barcode forward primer (F-5'-GCTTGGTTAGAATGGGTAACTAGTTTGCAG-3') and barcode-specific reverse primers for A1 (R-5'-TTCCCGAGCGCACCACAAA-3'), A3 (R-5'-ACACATACATCTCGCACGCAAACG-3'), B6 (R-5'-AAACCAACATCTCCCTCGCCC-3'), C2 (R-5'-GGTGATGTTTGGGCGTGGTAGAA-3'), C5 (R-5'-ATAACACTCCCGCCCACGAGAA-3'), D2 (R-5'-TTCGAACTCGACCGCCAACCAAA-3'), and D5 (R-5'-CCACTCAATCACGCAACACCC-3') with *E. coli* 16S primers (F-5'-ATTGACGTTACCCGCAGAAGA-3', R-5'-GGGATTTCACATCCGACTTGA-3') [28-29] on a QuantStudio 6 Real-Time PCR System. C_t values were normalized to *E. coli* 16S rRNA to

generate ΔC_t values, and fold changes were calculated by $\Delta\Delta C_t$ to the ΔC_t of the mouse inoculum.

Mock tissue-associated microbiome

Aliquots of purified genomic DNA isolated from independent strain cultures were mixed at equal abundance by weight. HMW DNA was isolated from a frozen aliquot of 10^7 HEK293 cells using the Circulomics Nanobind kit (https://www.circulomics.com/nanobind) per the manufacturer's recommended protocol. Human DNA was mixed with microbial mixture at a ratio of 99.5 : 0.5% to represent a realistic tissue-associated microbiota composition.

Full-length rRNA amplicon sequencing

For stool and *in vitro* host/microbiota samples, primers for proximal 16S (27F: AGRGTTTGATYHTGGCTCAG) and distal 23S (2241R: ACCRCCCCAGTHAAACT) were used to amplify the full-length rRNA operon (4,500bp). Starting with ~5ng expected microbial DNA, 25ul PCR reactions were prepared with LongAmp Taq 2x Master Mix (New England Biolabs, Inc; M0287L) and 0.4uM of each primer. We ran 20 cycles consisting of denaturation at 94C for 10s, annealing at 51C for 30s, and extension at 65C for 225s, followed by final extension at 65C for 10min. Typical yield is 700ng of full-length amplicons, which were verified by agarose gel electrophoresis (Fig S1) and fluorescent quantification (Qubit). Amplicon libraries were prepared for nanopore sequencing using the Oxford Nanopore Ligation Sequencing Kit (LSK109) per the manufacturer's protocol and sequenced on R9.4.1 flow cell on GridION with real-time basecalling using Guppy v3.2.2 in high-accuracy mode.

Results

Genome assembly and annotation

To establish a baseline for variation among known enteric *E. coli*, we produced finished genomes for eight AIEC and non-AIEC strains (NC101 [3,10], JA0018/A1, JA0019/A3, JA0022/B6, JA0036/C5, JA0044/D2, JA0048/D5, and JA0091/C2). These genomes were sequenced, assembled, polished, and annotated as described in the Methods. Each genome was assembled into a single complete, circular contig representing the entire genome (Table 2). Table 3 describes the serotype (O and H), multi-locus sequence type (MLST), and virulence and antibiotic genes found in each. Figure 2 illustrates these eight genomes, GC content, annotated genes, and approximate homology between them.

Strain	Genome size	Genes	GC%
JA0091/C2	5,141,716	6,129	50.56
JA0018/A1	5,124,838	6,451	50.67
JA0019/A3	4,860,314	5,759	50.72
JA0022/B6	4,775,206	5,693	50.78
JA0036/C5	5,025,403	6,100	50.86
JA0044/D2	5,019,749	7,083	50.58
JA0048/D5	4,906,020	5,822	50.55
NC101	5,029,635	6,128	50.76

Table 2. Assembly statistics for long-read E. coli genomes.

Strain	Serotype	MLST	Virulence genes	Antibiotic resistance
JA0091/C2	O2:H1	73	gad*,iss,pic*,ireA*,iha*,vat*,sat*	aph(3')-Ia*,blaEC-5*

JA0018/A1	O25:H1	73	gad*,mcmA,iss,mchB*,mchC*,cnf1*,mchF*,pic*,iroN*,iha*,vat*,sat*	aph(3')-Ia*,blaEC-5*
JA0019/A3	O4:H5	12	iss,gad*,vat*	aph(3')-Ia*,blaEC-5*
JA0022/B6	O83:H1	135	iss,gad*,vat*	aph(3')-Ia*,blaEC-19*
JA0036/C5	O25:H4	131	iha*,iss,gad*,sat*	aph(3')-Ia*,blaEC-19*
JA0044/D2	O16:H6	144	iha*,gad*,vat*,sat*	aph(3')-Ia*,blaEC-19*
JA0048/D5	O99:H6	1859	mchC*,mchB*,pic*,mchF*,iroN*,gad*,vat*	aph(3')-Ia*,blaEC-5*
NC101	O2:H6	998	sfaS*,iss,pic*,iroN*,gad*,vat*	blaEC-19*

Table 3. Characterization of each strain by serotype and MLST, including virulence, and antibiotic resistance genes.

Fig 2. Assembled genomes, illustrating relative GC content (red) and gene content (blue) for each strain, and approximate synteny between assemblies represented in gray, where darker bars represent higher identity (darkest gray is ~100% identity, white is <70%).

Strain-level variation

As illustrated in Figure 2, we observe significant genome-wide structural and nucleotide variation, even among these closely-related strains with similar adherent-invasive (AI) or non-AI phenotype. We computed the pairwise hamming distance between sequences across each strain and each copy of the rRNA operon for the 16S, ITS, 23S, and entire rRNA region (Figure 3).

Fig 3. Difference between 16S (a), 23S (b), ITS (c), and entire rRNA (d) sequences across strains and copies. Heatmap shows the relative number of discriminating sites (alleles) between

sequences. Sequences are clustered hierarchically using Ward's variance minimization and are labeled by strain (where K is K12, NC is NC101, and numerical labels are JA0XX strains).

We evaluated several regions independently for informative variation among these strains *in silico* by aligning our whole-genome sequence to either 1) the V1-V2 hypervariable region of 16S, 2) V3-V4 hypervariable region, 3) the entire 16S sequence containing all nine hypervariable regions and conserved spacers, 4) the entire rRNA operon, including 16S, ITS, and 23S, and 5) the entire genome. Very long nanopore reads allow us to use this as a proxy for PCR amplification of the various rRNA amplicon analyses since these reads average 10-20 Kbp (see Table 4), and thereby most often cover the entire region of interest.

Strain	# reads	Total Gbp	Average read length	N50
JA0091/C2	129,071	3.26	25,252	53,682
JA0018/A1	33,058	1.30	39,369	76,514
JA0019/A3	94,599	1.46	15,384	54,555
JA0022/B6	126,340	3.19	25,252	55,067
JA0036/C5	239,866	3.73	15,547	48,832
JA0044/D2	61,344	2.14	34,912	58,392
JA0048/D5	112,086	2.90	25,871	56,055
NC101	22,886	0.44	19,341	78,877

Table 4. Nanopore sequencing statistics for each strain, after demultiplexing.

In general, and not unexpectedly, the larger the region used for classification, the greater the accuracy in detecting the correct strain (Table 5). Hypervariable regions V1-V2 (27F-338R) and V3-V4 (343F-806R) produces poor results (averaging 35% and 27% accuracy, respectively),

little better than chance since there is little or no discriminating variation within that region. The entire 16S gene performs better (70%). Notably, there is a huge amount of variation in the accuracy across strains owing to their relative dissimilarity in a particular region. Full-length rRNA achieves an average 87.3% accuracy at the strain level, followed closely by all reads aligned against the entire genome (i.e. shotgun metagenomics) at 91%. Additional benefits of whole-metagenome sequencing include the observation of genes that may imply the functional capacity of the community without explicitly characterizing its taxonomic structure, and the ability to assemble the metagenome into an approximation of its component microbial genomes. These are important, tangible benefits, but are beyond the scope of this paper.

Strain	**V1-V2**	**V3-V4**	**16S**	**rRNA**	**Genomic**
JA0091/C2	14.48	29.66	61.95	71.75	74.85
JA0018/A1	14.72	12.50	47.39	70.73	84.79
JA0019/A3	17.52	19.61	96.93	93.44	90.44
JA0022/B6	93.18	18.66	97.62	87.07	94.78
JA0036/C5	89.62	87.39	99.06	99.19	97.09
JA0044/D2	14.68	31.25	59.16	96.83	97.57
JA0048/D5	17.60	2.90	55.26	99.07	96.13
NC101	16.67	10.77	45.71	80.18	93.43
Average	*34.81*	*26.59*	*70.39*	*87.28*	*91.14*

Table 5. Classification accuracy using commonly amplified 16S hypervariable regions, V1-V2 and V3-V4, the entire 16S and rRNA operons, and shotgun sequencing of the entire genome. While short hypervariable regions perform poorly and inconsistently across these samples, classification based on the entire rRNA operon (~4,500bp) approaches classification accuracy of whole-genome sequencing.

There is a relatively small improvement in classification accuracy using the entire genomes compared to the full-length rRNA operon. Like commonly used 16S primers that target subsets of the hypervariable regions, the primer pair used for the full operon are well-conserved across bacteria, but presents a much larger sequence with which to accurately classify sequences. This coupled with cost-effective multiplexing and long-read sequencing, despite much higher error rates than standard Illumina sequencing, should make this a viable approach for characterizing complex microbiome samples.

In vivo colonization studies

To evaluate the utility of our nanopore pipeline for complex biological samples, we performed *in vivo* colonization studies using a well-established IBD mouse model [3, 10-11, 30-32]. We first determined if these human-derived strains and murine-derived NC101 could colonize and persist in mice, in the presence of a competing microbiota. We colonized 2 singly-housed adult germ-free *Il10*$^{-/-}$ mice each with a single *E. coli* strain, then after one week provided niche competition by gavaging with a murine fecal transplant as outlined in Figure S2. After five weeks total, gastrointestinal tissues, including ileal tissue, cecal content, colon content, colon tissue, and colon mucus layer, were harvested and viable *E. coli* were quantified by serial plating. Table S1 reveals that almost all *E. coli* strains could colonize the lumen and intestinal tissue with complex community competition and no kanamycin to suppress the competing microbiota. Two strains of *E. coli* colonized inconsistently across both mice, with high levels of *E. coli* in one mouse and low or undetected levels of *E. coli* in the other. To verify if this would persist across a larger cohort, consistent colonizing strain, JA0048/D5, and inconsistent colonizing strain, JA0091/C2, were given by gavage to 7 or 8 germ-free *Il10*$^{-/-}$ mice, respectively. After one week we provided

niche competition by gavaging with a murine fecal transplant, then after three weeks we provided kanamycin water for 2 weeks to supress the competing microbiota as outlined in Figure S3. After ten weeks total, gastrointestinal tissues, including colon content, colon tissue, and colon mucus layer, were harvested and viable *E. coli* were quantified by serial plating. Table S2 reveals that consistent colonizing strain JA0048/D5 colonized all seven mice to high levels, while inconsistent colonizing strain JA0091/C2 colonized five of eight mice at the lumen and intestinal tissue.

To validate the pro-inflammatory potential of these strains, each strain was co-cultured with J774 macrophages and inflammatory colitis-inducing cytokine transcription and secretion [33-36]. The results in Figure 4 demonstrate that all strains induced inflammatory cytokine production.

Figure 4: Inflammatory cytokine production by macrophages stimulated with clinical E. coli *strains. (a-c) Relative mRNA abundance of each cytokine from stimulated J774 macrophages determined by qRT- PCR. Values normalized to gapdh, and fold increase relative to unstimulated controls. Each bar represents the mean across 3 independent experiments in triplicate and error bars show standard deviation (*p<0.05, Tukey's multiple comparisons test). If not otherwise noted, there is no significance. (d) Protein abundance of IL12(p40) from stimulated J774 macrophages determined by ELISA. Values normalized to NC101 stimulated macrophages. Each bar represents the mean across 4 independent experiments in duplicate and error bars show standard deviation (*p<0.05, Tukey's multiple comparisons test). If not otherwise noted, there is no significance*

To generate an *in vivo* microbiome containing varying amounts of the seven clinical *E. coli* strains and a complex community, we again colonized germ-free *Il10*$^{-/-}$ mice. Three mice were colonized by single oral gavage containing equal proportions of the seven clinical *E. coli* strains, and then housed in an SPF facility, where they become colonized over time with a natural mouse microbiome [3, 30]. After 10 weeks, the mice were sacrificed and stool samples were collected for nanopore sequencing and quantification of each strain by targeted qPCR.

Full-length rRNA metataxonomics

To further evaluate the utility of performing strain-level metataxonomics using full-length rRNA sequencing on an Oxford Nanopore device, we prepared and sequenced one stool sample from *in vivo Il10*$^{-/-}$ colonization studies depicted in Figure 5. Full-length rRNA amplification and sequencing using universal 16S 27F and 23S 2241R primers (see Methods) produced >600,000 reads with a median length of 4,149 bp (Table 6, Figure 6). We assigned these reads to individual rRNA copies by aligning to our database of known sequences in the assembled genomes (40 copies, five in each of eight strains). The proportion of reads assigned to each operon and strain is shown in Figure 5. There is a skewed representation of strains in the stool sample due to natural variation in their ability to colonize and thrive in the *Il10*$^{-/-}$ mouse gut. We confirmed the relative abundance of each strain in the stool sample through qPCR of the molecular barcode inserted in the neutral Tn7 site of each *E. coli* genome [28]. The qPCR showed a similar result, as there was skewing among the *E. coli* isolates in the mouse gut as shown in Figure 5 with the JA0036/C5 strain most abundant. As illustrated in Figure 7, there is also a large fraction of reads from stool sequencing (~75%) that originate from non-*Escherichia* genera and phyla as the mouse was housed under SPF, not germ-free conditions post-infection.

Sample	# reads	Total bp	Average read length	Median read length	Read N50
D stool M2	657,234	2.78 Gbp	4,229.65	4,149	4,194
0.5% *E.coli* mix, 99.5% HEK293	240,691	0.92 Gbp	3,829.50	4,136	4,173

Table 6. Sequencing statistics for full-length rRNA amplicon sequencing.

Figure 5. Strain distribution of seven E. coli isolates in murine stool sample. (a) Timeline for mouse experiment. (b) Relative percent abundance of each strain determined by full-length rRNA sequencing. Off-target hits to NC101 (4%) are expected to happen by chance given the classification error rate for the whole rRNA sequence (~13%, see Table 5). (c) Relative percent abundance of each strain determined by quantitative PCR of molecular barcode. Values normalized to E. coli *16S rRNA, and fold increase relative to pooled* E. coli *inoculum. Each bar represents the mean across 3 murine stool samples in duplicate and error bars show standard deviation.*

Figure 6. Read length distribution for stool and mock mixture full-length rRNA amplicon sequencing.

Figure 7. Distribution of nucleotide matches of the best alignment from amplified rRNA sequencing. Match/Mismatch represent the aligned rRNA segments from the barcoded whole-genome sequencing where we know the ground truth. Mismatches have a nearly identical distribution to matches, indicating sequencing errors caused miscategorization as a very closely-

related rRNA. Mock shows a very similar distribution, but sequences from stool are bimodal indicating a (smaller) proportion of reads originated from these or similar E.coli rRNA, and a majority (~75%) originating from highly divergent (70-80% identity) likely originating from different genera and phyla.

We additionally prepared a mock mixture consisting of all eight strains at even abundance mixed 0.5 : 99.5% with human DNA (HEK293) to simulate a realistic tissue-associated microbiota sample. Full-length rRNA sequences were successfully amplified and sequenced as done with the stool sample. We observed relatively uniform representation of the eight strains in the mock mixture, demonstrating our ability to amplify and properly classify closely-related rRNA sequences in a host-tissue-like context, as illustrated in Figure 8.

Figure 8. Strain representation based on full-length rRNA sequencing from mock host:microbial mixture. Mock mixture contained an approximately uniform quantify of all eight strains before amplification.

Discussion

Dysbiosis of gut microbial communities are implicated in both IBD and CRC, with an increase of mucosally-adhered bacteria. *E. coli* are found in greater proportion among IBD patients, including AIEC strains that are determined through the *in vitro* ability to adhere to and invade epithelial cells and survive and replicate in macrophages. AIEC is a functional definition, and it is increasingly important to compare this *in vitro* behavior with colonization ability at the mucosa in the gut. The clearer our definition of who is colonizing intestinal tissues, the closer

we will get to defining molecular features that are common among AIEC that could predict tissue colonization. To be able to do personalized medicine, we need to be able to distinguish individual strains from within a complex community. This is especially important in biopsies of IBD patients, looking at active disease lesions vs. adjacent normal tissue, and even on tumor/off tumor in colorectal cancer samples.

This focused study highlights the challenges in resolving closely-related but functionally divergent strains within mixed microbial populations. The commonly used metataxonomic approach of sequencing one or two amplified hypervariable regions of the 16S gene are unable to distinguish among these strains. However, we demonstrate that there exists substantial genetic and genomic variation between them that can be detected using long-read sequencing. We illustrate the difference in sensitivity and specificity of strain detection using full-length rRNA amplification and whole-metagenome sequencing.

While these results suggest a method for sensitive identification of known strains, this approach does not suggest a general taxonomic analysis protocol for microbiome studies since the genomes of the strains of interest must be known *a priori* in order to properly identify them. Full-length rRNA sequencing following PCR amplification permits accurate strain identification even when the microbial abundance - for example in tissue-associated microbial samples - is very low. We demonstrated high concordance among relative abundance between sequenced full-length rRNA sequences and qPCR results targeting the inserted molecular barcodes, suggesting that abundances are reliably preserved through amplification and nanopore sequencing. However, even higher accuracy and explicit identification of microbial genic content can be captured by shotgun metagenome sequencing when there exists enough material to extract high-quality microbial DNA.

Future work includes optimization of the full-length rRNA protocol for tissue samples with adherent microbes. Our analysis of *in vitro* mixtures indicate this should be feasible with microbial:host DNA content as low as 0.5% - consistent with expected abundances [37-38] in many human mucosa, including the gut - but we have so far failed to amplify full-length rRNA from resected mouse gut. Additionally, given the extraordinary utility of shotgun metagenomic sequence, it is worthwhile to explore the several existing methods for selective extraction and sequencing of microbial DNA to expand the availability of non-PCR-based approaches in the context of tissue-associated microbial communities.

Acknowledgements

This work was supported by the following grants: NIH/NIDDK K01 DK119582 (JW), SPIRE (NIH IRACDA program) NIH/NIGMS 5K12 GM000678-21 (RMB), NIH/NIDDK K01 DK103952 (JCA), American Gastroenterological Association Augustyn Award in Digestive Cancer (JCA), Lineberger Comprehensive Cancer Center Pilot Grant and UCRF (JCA).

We acknowledge the Gnotobiotic Core at the UNC Center for Gastrointestinal Biology and Disease (CGIBD: supported by NIH P30DK34987), the National Gnotobiotic Rodent Resource Center (supported by NIH P40 OD01995), and UNC's High-Throughput Sequencing Facility. We also acknowledge Christopher Broberg, Adrienne Franks, and Cassandra Barlogio for experimental assistance.

References

1. Sartor RB, Wu GD. Roles for Intestinal Bacteria, Viruses, and Fungi in Pathogenesis of Inflammatory Bowel Diseases and Therapeutic Approaches. Gastroenterology. 2017;152(2):327-339.e4.

2. Kappelman MD, Farkas DK, Long MD, Erichsen R, Sandler RS, Sørensen HT, et al. Risk of cancer in patients with inflammatory bowel diseases: A nationwide population-based cohort study with 30 years of follow-up evaluation. Clin Gastroenterol Hepatol [Internet]. 2014;12(2):265-273.e1. Available from: http://dx.doi.org/10.1016/j.cgh.2013.03.034
3. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan T-J. Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota. Science 05 Oct 2012: Vol. 338, Issue 6103, pp. 120-123.
4. Dejea CM, Fathi P, Craig JM, Boleij A, Taddese R, Geis AL, et al. Patients with familial adenomatous polyposis harbor colonic biofilms containing tumorigenic bacteria. Science (80-). 2018;359(6375):592–7.
5. Darfeuille-Michaud A, Boudeau J, Bulois P, Neut C, Glasser A-L, Barnich N et al. High Prevalence of Adherent-Invasive Escherichia Coli Associated With Ileal Mucosa in Crohn's Disease. Gastroenterology, 127 (2), 412-21 Aug 2004.
6. Martin HM, Campbell BJ, Hart CA, Mpofu C, Nayar M, Singh R, et al. Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer. Gastroenterology. 2004;127(1):80–93.
7. Baumgart M, Dogan B, Rishniw M, Weitzman G, Bosworth B, Yantiss R, et al. Culture independent analysis of ileal mucosa reveals a selective increase in invasive Escherichia coli of novel phylogeny relative to depletion of Clostridiales in Crohn's disease involving the ileum. ISME J. 2007;1(5):403–18.
8. Knights D, Lassen KG, Xavier RJ. Advances in inflammatory bowel disease pathogenesis: Linking host genetics and the microbiome. Gut. 2013;62(10):1505–10.
9. Frank DN, St. Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci U S A. 2007;104(34):13780–5.
10. Kim SC, Tonkonogy SL, Albright CA, Tsang J, Balish EJ, Braun J, et al. Variable phenotypes of enterocolitis in interleukin 10-deficient mice monoassociated with two different commensal bacteria. Gastroenterology. 2005;128(4):891–906.
11. Kim SC, Tonkonogy SL, Karrasch T, Jobin C, Balfour Sartor R. Dual-association of gnotobiotic IL-10-/- mice with 2 nonpathogenic commensal bacteria induces aggressive pancolitis. Inflamm Bowel Dis. 2007;13(12):1457–66.
12. Carvalho FA, Koren O, Goodrich JK, Johansson MEV, Nalbantoglu I, Aitken JD, et al. Transient inability to manage proteobacteria promotes chronic gut inflammation in TLR5-deficient mice. Cell Host Microbe. 2012;12(2):139–52.
13. O'Brien CL, Bringer MA, Holt KE, Gordon DM, Dubois AL, Barnich N, et al. Comparative genomics of Crohn's disease-Associated adherent-invasive Escherichia coli. Gut. 2017;66(8):1382–9.
14. Benítez-Páez A, Portune KJ, Sanz Y. Species-level resolution of 16S rRNA gene amplicons sequenced through the MinION™ portable nanopore sequencer. GigaScience, Volume 5, Issue 1, December 2016, s13742-016-0111-z.
15. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon [version 2; peer review: 2 approved, 3 approved with reservations]. F1000Research 2019, 7:1755.

16. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS et al. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Research, Volume 47, Issue 18, 10 October 2019, Page e103.
17. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q et al. Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. https://www.biorxiv.org/content/10.1101/645903v3. 2019.
18. Sambrook J, Russell DW. Purification of Nucleic Acids by Extraction With Phenol:chloroform. CSH Protoc, 2006 (1).
19. Tyson J. Rocky Mountain adventures in Genomic DNA sample preparation, ligation protocol optimisation / simplification and Ultra long read generation. protocols.io. 2020. dx.doi.org/10.17504/protocols.io.7euhjew.
20. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics, Volume 32, Issue 14, 15 July 2016, Pages 2103–2110,
21. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. gr.214270.116Published in Advance January 18, 2017, doi:10.1101/gr.214270.116.
22. Oxford Nanopore Technologies. Medaka; 2018 [cited 2020 March 3]. Available from: https://nanoporetech.github.io/medaka/
23. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016 Aug 19;44(14):6614-24. doi: 10.1093/nar/gkw569.
24. Buchfink B, Xie C. Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods volume 12, pages59–60(2015).
25. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014; 6(11): 90.
26. Center for Genomic Epidemiology; 2011 [cited 2020 March 3]. [Internet]. Available from: http://www.genomicepidemiology.org/
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, Volume 34, Issue 18, 15 September 2018, Pages 3094–3100.
28. Walters MS, Lane MC, Vigil PD, Smith SN, Walk SR, Mobley HLT. Kinetics of Uropathogenic Escherichia Coli Metapopulation Movement During Urinary Tract Infection. mBio, 3 (1) 2012 Feb 7,
29. Gonzalez RJ, Lane MC, Wagner NJ, Weening EH, Miller VL. Dissemination of a highly virulent pathogen: tracking the early events that define infection. PLoS Pathog. 2015;11(1):e1004587.
30. Arthur JC, Gharaibeh RZ, Mühlbauer M, Perez-Chanona E, Uronis JM, McCafferty J et al. Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer. Nature Communications volume 5, Article number: 4724 (2014).
31. Kühn R, Löhler J, Rennick D, Rajewsky K, Müller W. Interleukin-10-deficient mice develop chronic enterocolitis. Cell. 1993;75(2):263–74.
32. Ellermann M, Gharaibeh RZ, Fulbright L, Dogan B, Moore LN, Broberg CA et al. Yersiniabactin-Producing Adherent/Invasive *Escherichia coli* Promotes Inflammation-

Associated Fibrosis in Gnotobiotic *Il10*$^{-/-}$ Mice. Infection and Immunity. 2019;87(11):1–18.

33. Sartor RB. Cytokines in intestinal inflammation: Pathophysiological and clinical considerations. Gastroenterology. 1994;106(2):533–9.
34. Patwa LG, Fan TJ, Tchaptchet S, Liu Y, Lussier YA, Sartor RB, et al. Chronic intestinal inflammation induces stress-response genes in commensal Escherichia coli. Gastroenterology [Internet]. 2011;141(5):1842-1851.e10. Available from: http://dx.doi.org/10.1053/j.gastro.2011.06.064
35. Ellermann M, Huh EY, Liu B, Carroll IM, Tamayo R, Sartor RB. Adherent-invasive Escherichia coli production of cellulose influences iron-induced bacterial aggregation, phagocytosis, and induction of colitis. Infect Immun. 2015;83(10):4068–80.
36. Segain JP, Galmiche JP, Raingeard De La Blétière D, Bourreille A, Leray V, Gervois N, et al. Butyrate inhibits inflammatory responses through NFκB inhibition: Implications for Crohn's disease. Gut. 2000;47(3):397–403.
37. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012, 486 207–214. 10.1038/nature11234.
38. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et alStrains, functions and dynamics in the expanded human microbiome project. Nature 2017, 550 61–66. 10.1038/nature23889.

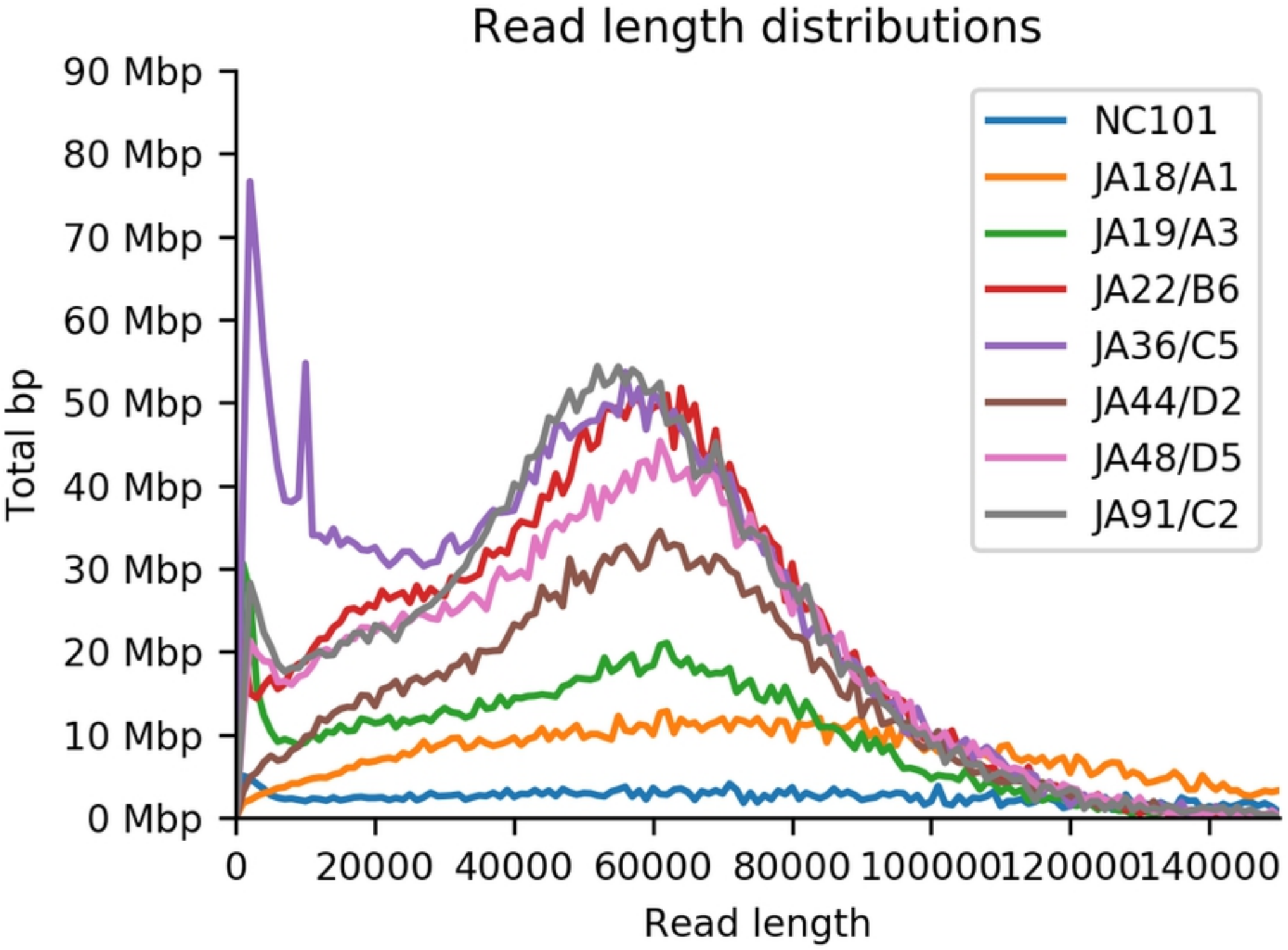
Read length distributions
90 Mbp
80 Mbp
70 Mbp
60 Mbp
50 Mbp
40 Mbp
30 Mbp
20 Mbp
10 Mbp
0 Mbp
Total bp
0
20000
40000
60000
80000
100000
120000
140000
Read length
NC101
JA18/A1
JA19/A3
JA22/B6
JA36/C5
JA44/D2
JA48/D5
JA91/C2

Fig 1

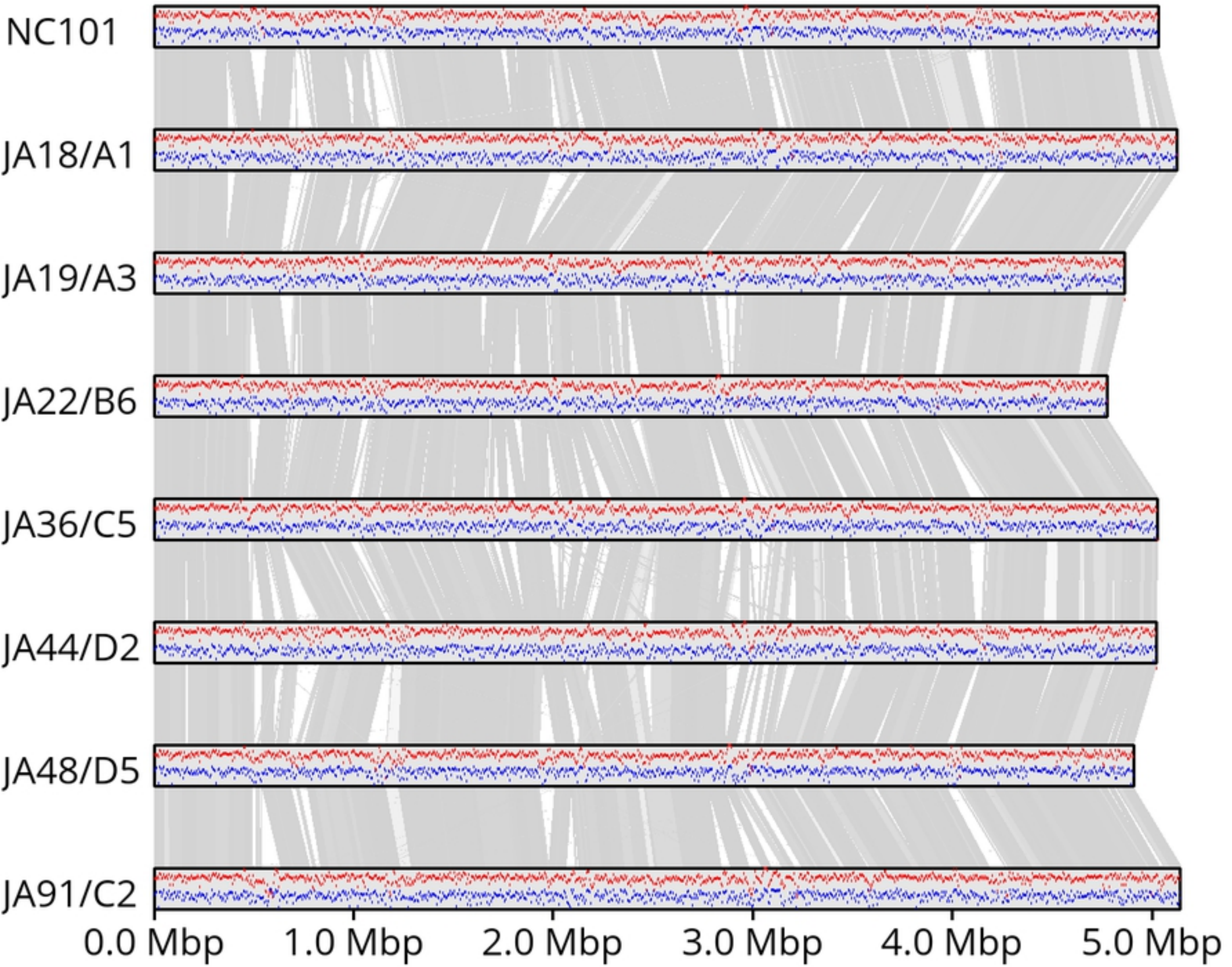

Fig 2

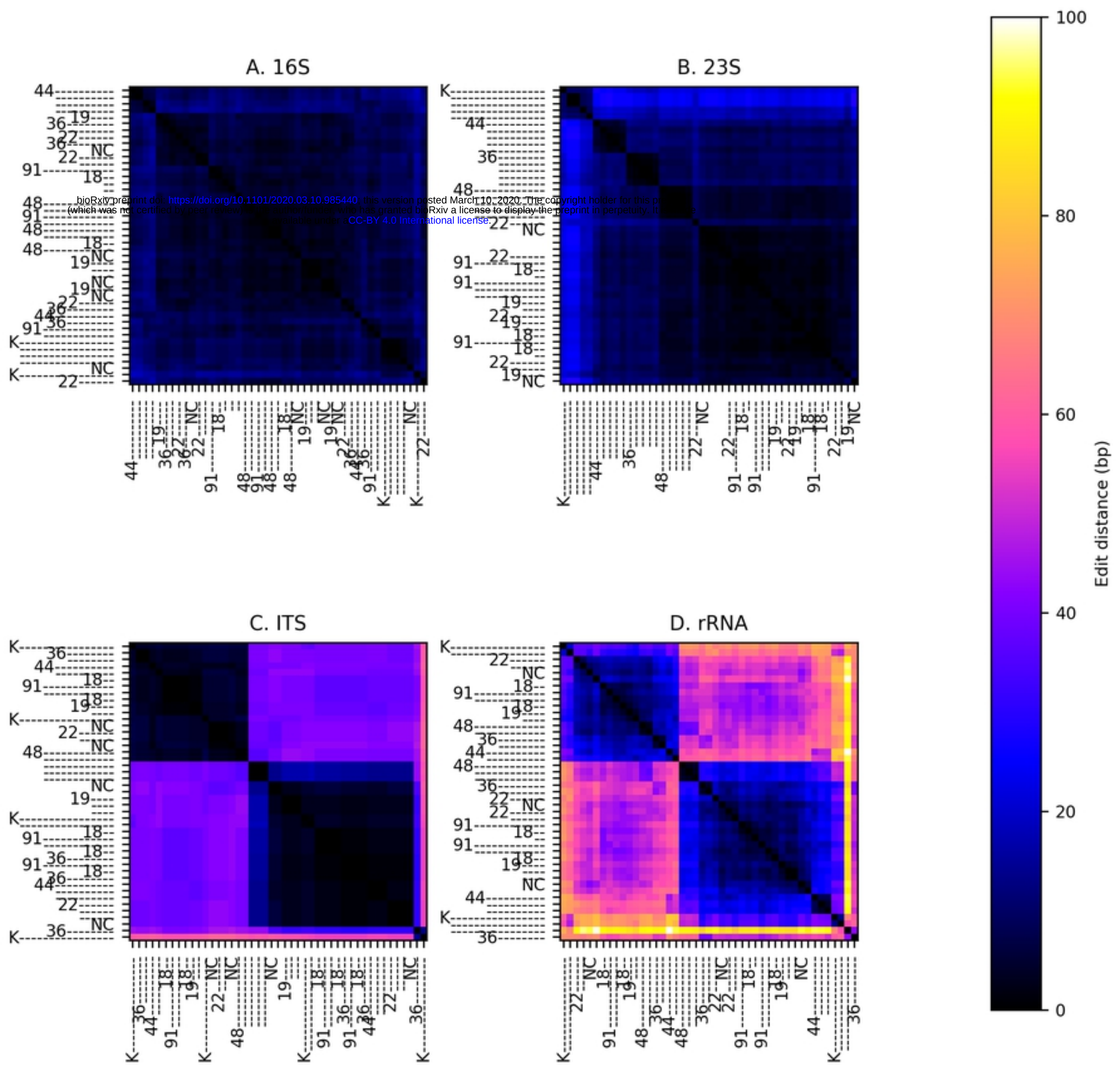

Fig 3

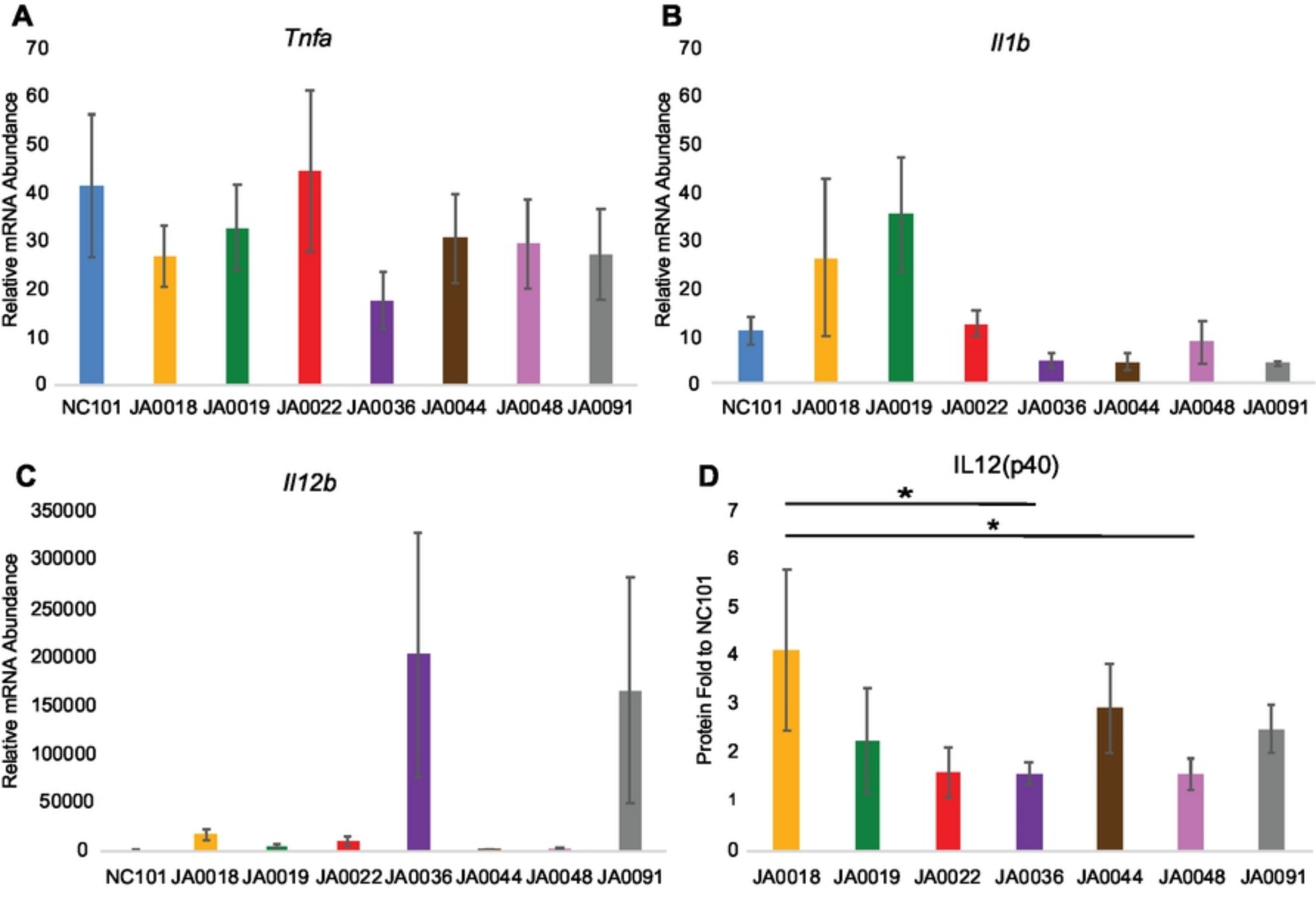
A
Tnfa
Relative mRNA Abundance
70
60
50
40
30
20
10
0
NC101 JA0018 JA0019 JA0022 JA0036 JA0044 JA0048 JA0091
B
Il1b
Relative mRNA Abundance
70
60
50
40
30
20
10
0
NC101 JA0018 JA0019 JA0022 JA0036 JA0044 JA0048 JA0091
C
Il12b
Relative mRNA Abundance
350000
300000
250000
200000
150000
100000
50000
0
NC101 JA0018 JA0019 JA0022 JA0036 JA0044 JA0048 JA0091
D
IL12(p40)
*
*
Protein Fold to NC101
7
6
5
4
3
2
1
0
JA0018 JA0019 JA0022 JA0036 JA0044 JA0048 JA0091

Fig 4

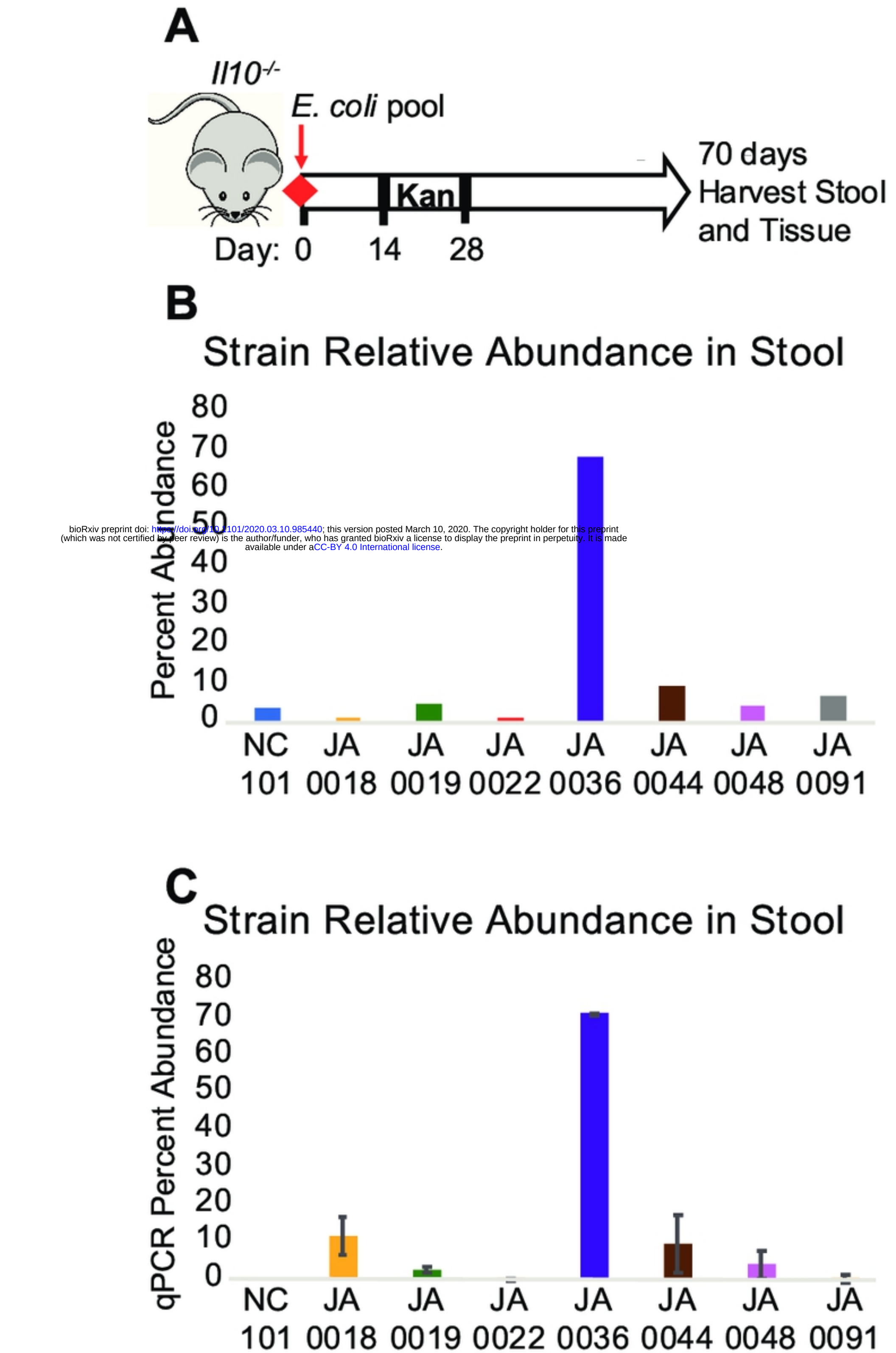

Fig 5

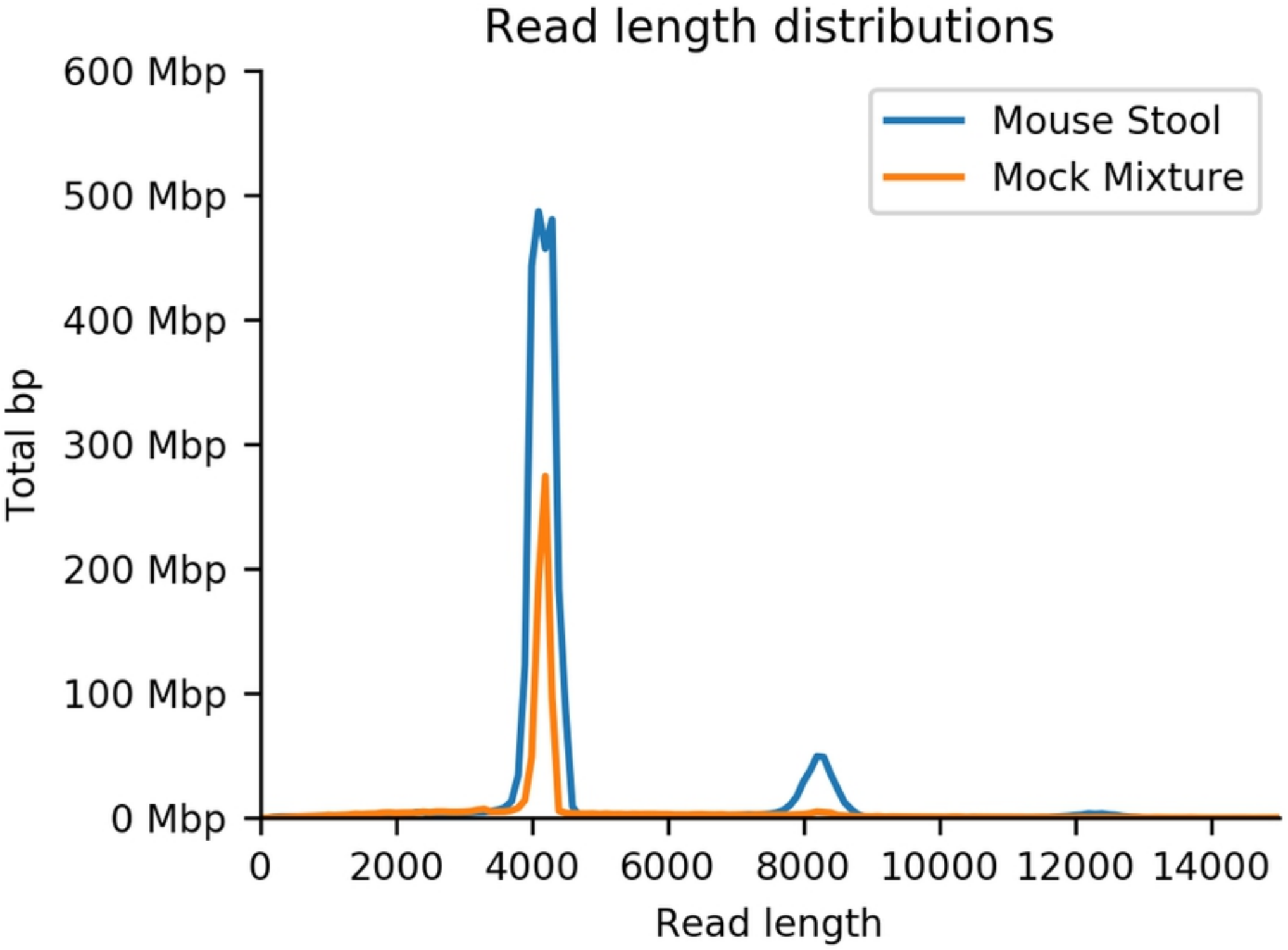
Read length distributions
600 Mbp
500 Mbp
400 Mbp
300 Mbp
200 Mbp
100 Mbp
0 Mbp
Total bp
0
2000
4000
6000
8000
10000
12000
14000
Read length
Mouse Stool
Mock Mixture

Fig 6

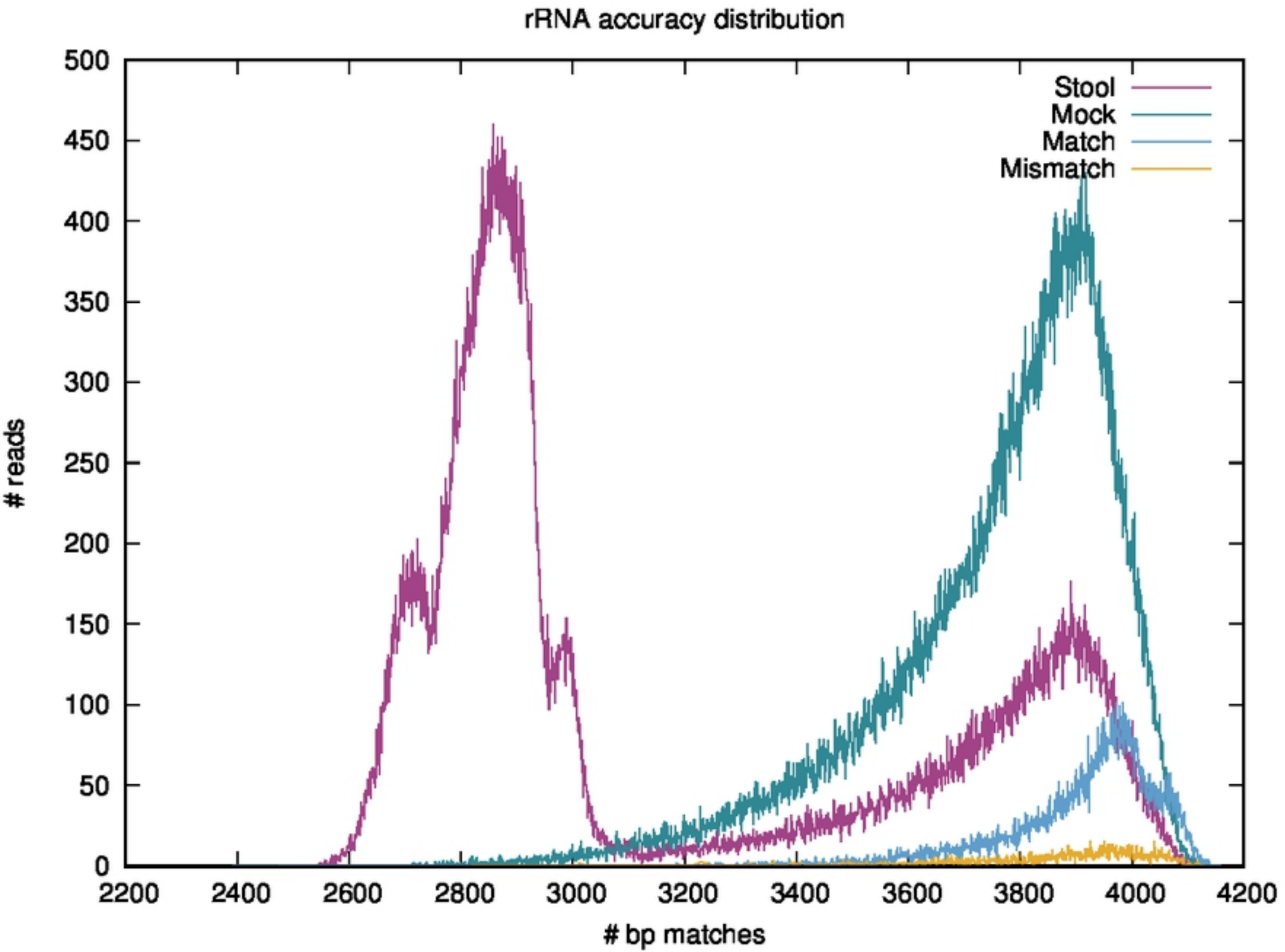
rRNA accuracy distribution
Stool
Mock
Match
Mismatch
reads
bp matches
0
50
100
150
200
250
300
350
400
450
500
2200
2400
2600
2800
3000
3200
3400
3600
3800
4000
4200

Fig 7

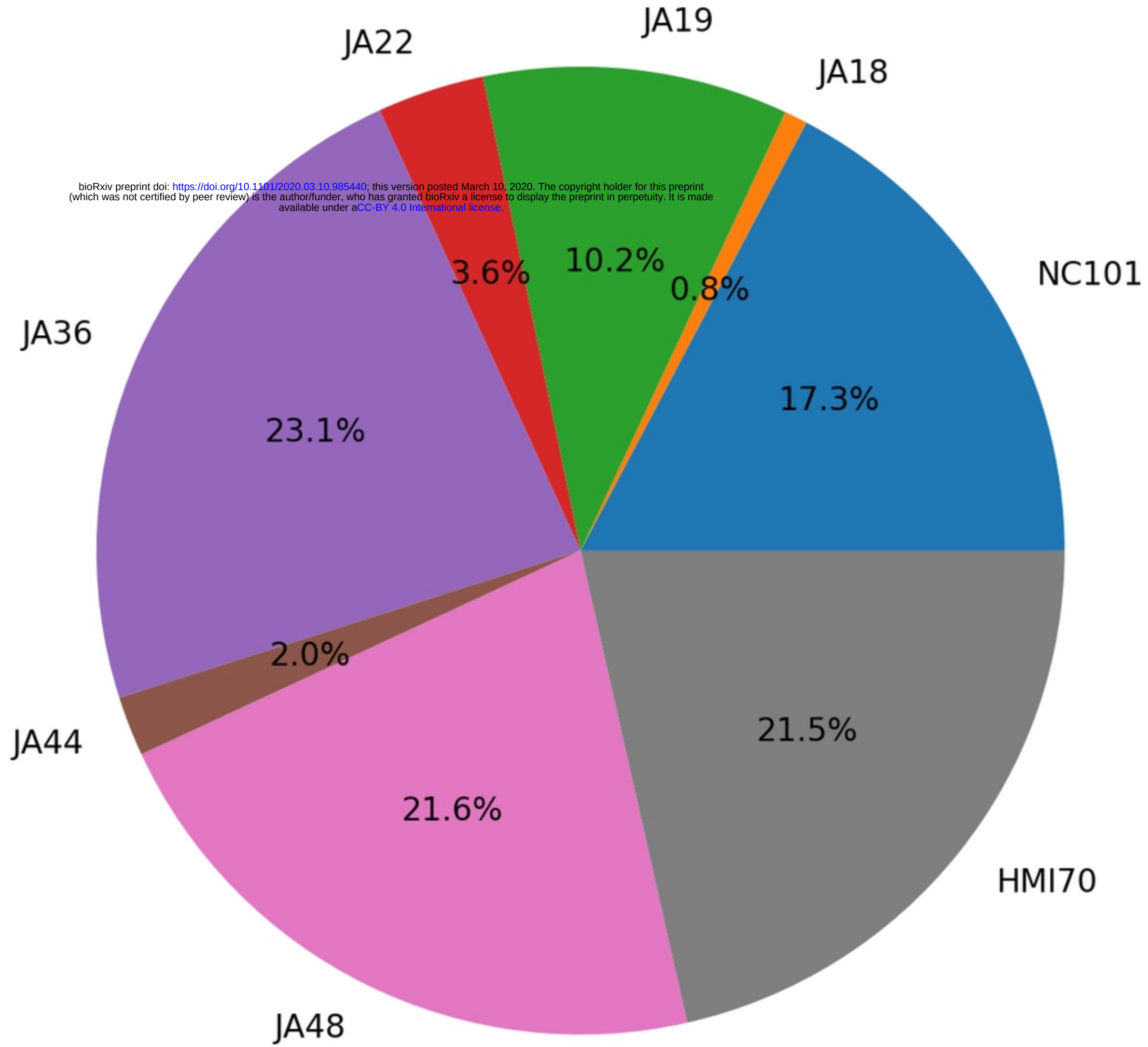

Fig 8