From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: NCBI check: contamination still present
Date: June 19, 2017 at 5:45 PM
To: Terrance Shea tshea@broadinstitute.org
Cc: Sarah Young stowey@broadinstitute.org, Jim Bochicchio jboch@broadinstitute.org, Aaron M Ellison aellison@fas.harvard.edu, Caroline Cusick ccusick@broadinstitute.org

Great, thanks Terry. I'll get those submitted to NCBI and get those checked.

Best,
Matt

> On Jun 19, 2017, at 5:36 PM, Terrance Shea <tshea@broadinstitute.org> wrote:
>
> Hi Matt, all-
>
> I apologize for missed adapter. We will have to take another look at our screening updates.
>
> I have removed the regions reported by NCBI for SM-AZXXN and SM-AZXXO.
>
> In the FTP area there is a new folder titled "20170619" and within this are two .tar.gz files for SM-AZXXN and SM-AZXXO.
>
> Let us know if anything further is found.
>
> Terry
>
> On Mon, Jun 19, 2017 at 11:56 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>> Thanks Sarah!
>>
>> Also, Terry, just got this too:
>>
>>
>> SUBID BioProject BioSample Organism ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster picea [] We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the

sequences appropriately and then resubmit your sequences. After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. [] Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly. Skipped 43,216 same as before; screening 18 sequences, 128,738 bp. Note: 4,997 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 19 sequences with locations to mask/trim (21 split spans with locations to mask/trim) Trim: Sequence name, length, span(s), apparent source scaffold00007 1562141 1022691..1022714 adaptor:NGB00749.1-not_cleaned scaffold00009 1451161 1122657..1122692 adaptor:NGB00843.1-not_cleaned scaffold00033 1132792 320706..320740,792366..792399 adaptor:NGB00843.1-not_cleaned scaffold00039 1045078 881378..881401 adaptor:NGB00843.1-not_cleaned scaffold00084 761745 375848..375881 adaptor:NGB00843.1-not_cleaned scaffold00093 733569 80388..80421 adaptor:NGB00843.1-not_cleaned scaffold00099 709502 254618..254654 adaptor:NGB00843.1-not_cleaned scaffold00114 634545 581610..581643 adaptor:NGB00843.1-not_cleaned scaffold00115 625950 229337..229381 adaptor:NGB00843.1-not_cleaned scaffold00126 590179 169372..169405 adaptor:NGB00843.1-not_cleaned scaffold00199 429751 396673..396706,406816..406849 adaptor:NGB00843.1-not_cleaned scaffold00281 328988 102353..102391 adaptor:NGB00843.1-not_cleaned scaffold00296 318450 116590..116623 adaptor:NGB00843.1-not_cleaned scaffold00435 217246 114637..114666 adaptor:NGB00843.1-not_cleaned scaffold00445 210205 94262..94293 adaptor:NGB00843.1-not_cleaned scaffold00547 157883 40880..40908 adaptor:NGB00843.1-not_cleaned scaffold00699

110410 44231..44263 adaptor:NGB00843.1-not_cleaned scaffold00850
83541 56728..56763 adaptor:NGB00843.1-not_cleaned scaffold05365
3477 1..44 adaptor:NGB00843.1-not_cleaned

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

On Jun 19, 2017, at 8:29 AM, Sarah Young <stowey@broadinstitute.org> wrote:

> Looping Terry in.
>
> On Sun, Jun 18, 2017 at 5:33 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>
> > Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).
> >
> > Can you guys remove these and check the other genomes ASAP?
> >
> > Thanks,
> >
> > Matt

Matt

SUBID BioProject BioSample Organism ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []

We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.

After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. []

Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly.

Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to mask/trim)

Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00843.1 scaffold00282 286384 154047..154082 adaptor:NGB00843.1 scaffold00424 169256

> > > 91684..91733 adaptor:NGB00843.1 scaffold00635 94765
> > > 70145..70194 adaptor:NGB00843.1 scaffold02309 19718
> > > 14150..14191 adaptor:NGB00843.1 scaffold03495 10772 8825..8858
> > > adaptor:NGB00843.1
> > >
> > > Matt
> > >
> > > Postdoctoral Research Fellow
> > > Harvard Forest
> > > Harvard University
> > > 324 N Main St
> > > Petersham, MA 01366
> > >
> > > Office: (978) 756-6165
> > >
> > > Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.
> > >
> > > -- Miles Kington

--
Sarah Young
Director, Operations, Finance and Computation
Broad Technology Labs
The Broad Institute
75 Ames Street
Cambridge, MA 02141
T: (617) 714-8508
E: stowey@broadinstitute.org

From: **Terrance Shea** tshea@broadinstitute.org
Subject: Re: NCBI check: contamination still present
Date: June 19, 2017 at 5:37 PM
To: Lau, Matthew K. matthewklau@fas.harvard.edu
Cc: Sarah Young stowey@broadinstitute.org, Jim Bochicchio jboch@broadinstitute.org, Ellison, Aaron aellison@fas.harvard.edu, Caroline Cusick ccusick@broadinstitute.org

TS

Hi Matt, all-

I apologize for missed adapter. We will have to take another look at our screening updates.

I have removed the regions reported by NCBI for SM-AZXXN and SM-AZXXO.

In the FTP area there is a new folder titled "20170619" and within this are two .tar.gz files for SM-AZXXN and SM-AZXXO.

Let us know if anything further is found.

Terry

On Mon, Jun 19, 2017 at 11:56 AM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:

> Thanks Sarah!
>
> Also, Terry, just got this too:
>
> SUBID BioProject BioSample Organism ---------- ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster picea [] We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences. After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing

platforms share sequence similarity. [] Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly. Skipped 43,216 same as before; screening 18 sequences, 128,738 bp. Note: 4,997 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 19 sequences with locations to mask/trim (21 split spans with locations to mask/trim) Trim: Sequence name, length, span(s), apparent source scaffold00007 1562141 1022691..1022714 adaptor:NGB00749.1-not_cleaned scaffold00009 1451161 1122657..1122692 adaptor:NGB00843.1-not_cleaned scaffold00033 1132792 320706..320740,792366..792399 adaptor:NGB00843.1-not_cleaned scaffold00039 1045078 881378..881401 adaptor:NGB00843.1-not_cleaned scaffold00084 761745 375848..375881 adaptor:NGB00843.1-not_cleaned scaffold00093 733569 80388..80421 adaptor:NGB00843.1-not_cleaned scaffold00099 709502 254618..254654 adaptor:NGB00843.1-not_cleaned scaffold00114 634545 581610..581643 adaptor:NGB00843.1-not_cleaned scaffold00115 625950 229337..229381 adaptor:NGB00843.1-not_cleaned scaffold00126 590179 169372..169405 adaptor:NGB00843.1-not_cleaned scaffold00199 429751 396673..396706,406816..406849 adaptor:NGB00843.1-not_cleaned scaffold00281 328988 102353..102391 adaptor:NGB00843.1-not_cleaned scaffold00296 318450 116590..116623 adaptor:NGB00843.1-not_cleaned scaffold00435 217246 114637..114666 adaptor:NGB00843.1-not_cleaned scaffold00445 210205 94262..94293 adaptor:NGB00843.1-not_cleaned scaffold00547 157883 40880..40908 adaptor:NGB00843.1-not_cleaned scaffold00699 110410 44231..44263 adaptor:NGB00843.1-not_cleaned scaffold00850 83541 56728..56763 adaptor:NGB00843.1-not_cleaned scaffold05365 3477 1..44 adaptor:NGB00843.1-not_cleaned

Matt

Postdoctoral Research Fellow
Harvard Forest

Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

On Jun 19, 2017, at 8:29 AM, Sarah Young <stowey@broadinstitute.org> wrote:

> Looping Terry in.
>
> On Sun, Jun 18, 2017 at 5:33 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>> Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).
>>
>> Can you guys remove these and check the other genomes ASAP?
>>
>> Thanks,
>>
>> Matt
>>
>>
>> SUBID BioProject BioSample Organism ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []
>>
>> We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.

After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. []

Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly.

Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to mask/trim)

Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00843.1 scaffold00282 286384 154047..154082 adaptor:NGB00843.1 scaffold00424 169256 91684..91733 adaptor:NGB00843.1 scaffold00635 94765 70145..70194 adaptor:NGB00843.1 scaffold02309 19718 14150..14191 adaptor:NGB00843.1 scaffold03495 10772 8825..8858 adaptor:NGB00843.1

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University

324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

--
Sarah Young
Director, Operations, Finance and Computation
Broad Technology Labs
The Broad Institute
75 Ames Street
Cambridge, MA 02141
T: (617) 714-8508
E: stowey@broadinstitute.org

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: NCBI check: contamination still present
Date: June 19, 2017 at 11:56 AM
To: Sarah Young stowey@broadinstitute.org
Cc: Terrance Shea tshea@broadinstitute.org, Jim Bochicchio jboch@broadinstitute.org, Ellison, Aaron aellison@fas.harvard.edu, Caroline Cusick ccusick@broadinstitute.org

Thanks Sarah!

Also, Terry, just got this too:

SUBID BioProject BioSample Organism ---------- SUB2631470 PRJNA385595 Aphaenogaster picea [] We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences. After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. [] Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly. Skipped 43,216 same as before; screening 18 sequences, 128,738 bp. Note: 4,997 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 19 sequences with locations to mask/trim (21 split spans with locations to mask/trim) Trim: Sequence name, length, span(s), apparent source scaffold00007 1562141 1022691..1022714 adaptor:NGB00749.1-not_cleaned scaffold00009 1451161 1122657..1122692 adaptor:NGB00843.1-not_cleaned scaffold00033 1132792 320706..320740,792366..792399 adaptor:NGB00843.1-not_cleaned scaffold00039 1045078 881378..881401 adaptor:NGB00843.1-not_cleaned scaffold00084 761745 375848..375881 adaptor:NGB00843.1-not_cleaned scaffold00093 733569 80388..80421 adaptor:NGB00843.1-not_cleaned scaffold00099 709502 254618..254654 adaptor:NGB00843.1-not_cleaned

scaffold00114 634545 581610..581643 adaptor:NGB00843.1-not_cleaned
scaffold00115 625950 229337..229381 adaptor:NGB00843.1-not_cleaned
scaffold00126 590179 169372..169405 adaptor:NGB00843.1-not_cleaned
scaffold00199 429751 396673..396706,406816..406849
adaptor:NGB00843.1-not_cleaned scaffold00281 328988 102353..102391
adaptor:NGB00843.1-not_cleaned scaffold00296 318450 116590..116623
adaptor:NGB00843.1-not_cleaned scaffold00435 217246 114637..114666
adaptor:NGB00843.1-not_cleaned scaffold00445 210205 94262..94293
adaptor:NGB00843.1-not_cleaned scaffold00547 157883 40880..40908
adaptor:NGB00843.1-not_cleaned scaffold00699 110410 44231..44263
adaptor:NGB00843.1-not_cleaned scaffold00850 83541 56728..56763
adaptor:NGB00843.1-not_cleaned scaffold05365 3477 1..44
adaptor:NGB00843.1-not_cleaned

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

On Jun 19, 2017, at 8:29 AM, Sarah Young <stowey@broadinstitute.org> wrote:

> Looping Terry in.
>
> On Sun, Jun 18, 2017 at 5:33 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:

<matthewklau@fas.harvard.edu> wrote:

Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).

Can you guys remove these and check the other genomes ASAP?

Thanks,

Matt

SUBID BioProject BioSample Organism ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []

We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.

After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. []

Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly.

Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with

> locations to mask/trim)
>
> Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00843.1 scaffold00282 286384 154047..154082 adaptor:NGB00843.1 scaffold00424 169256 91684..91733 adaptor:NGB00843.1 scaffold00635 94765 70145..70194 adaptor:NGB00843.1 scaffold02309 19718 14150..14191 adaptor:NGB00843.1 scaffold03495 10772 8825..8858 adaptor:NGB00843.1
>
> Matt
>
> Postdoctoral Research Fellow
> Harvard Forest
> Harvard University
> 324 N Main St
> Petersham, MA 01366
>
> Office: (978) 756-6165
>
> Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.
>
> -- Miles Kington

--
Sarah Young
Director, Operations, Finance and Computation
Broad Technology Labs
The Broad Institute
75 Ames Street
Cambridge, MA 02141
T: (617) 714-8508
E: stowey@broadinstitute.org

From: **Sarah Young** stowey@broadinstitute.org
Subject: Re: NCBI check: contamination still present
Date: June 19, 2017 at 8:29 AM
To: Lau, Matthew K. matthewklau@fas.harvard.edu, Terrance Shea tshea@broadinstitute.org
Cc: Jim Bochicchio jboch@broadinstitute.org, Ellison, Aaron aellison@fas.harvard.edu, Caroline Cusick ccusick@broadinstitute.org

SY

Looping Terry in.

On Sun, Jun 18, 2017 at 5:33 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:

> Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).
>
> Can you guys remove these and check the other genomes ASAP?
>
> Thanks,
>
> Matt
>
> SUBID BioProject BioSample Organism ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []
>
> We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.
>
> After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. []
>
> Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then

you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly.

Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to mask/trim)

Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00843.1 scaffold00282 286384 154047..154082 adaptor:NGB00843.1 scaffold00424 169256 91684..91733 adaptor:NGB00843.1 scaffold00635 94765 70145..70194 adaptor:NGB00843.1 scaffold02309 19718 14150..14191 adaptor:NGB00843.1 scaffold03495 10772 8825..8858 adaptor:NGB00843.1

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

--
Sarah Young

Sarah Young
Director, Operations, Finance and Computation
Broad Technology Labs
The Broad Institute
75 Ames Street
Cambridge, MA 02141
T: (617) 714-8508
E: stowey@broadinstitute.org

From: **Ellison, Aaron** aellison@fas.harvard.edu
Subject: RE: NCBI check: contamination still present
Date: June 18, 2017 at 6:06 PM
To: Lau, Matthew K. matthewklau@fas.harvard.edu

AE

I hope they do it quickly.

Best,
Aaron

From: Lau, Matthew K.
Sent: Sunday, June 18, 2017 18:02
To: Ellison, Aaron <aellison@fas.harvard.edu>
Subject: Re: NCBI check: contamination still present

Hey Aaron, they may need to trim the Ns and the end of the sequences too, I didn't check this, but the end of the error report lists the adapters that were detected and where in the genomes.

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

On Jun 18, 2017, at 5:43 PM, Ellison, Aaron <aellison@fas.harvard.edu> wrote:

> It looks like this is a consequence of too many "N"s. Is that the correct interpretation?
>
> Thanks for resolving this as quickly as possible,
> Aaron

From: Lau, Matthew K.
Sent: Sunday, June 18, 2017 17:33
To: Jim Bochicchio <jboch@broadinstitute.org>; Ellison, Aaron <aellison@fas.harvard.edu>; Sarah Young <stowey@broadinstitute.org>; Caroline Cusick <ccusick@broadinstitute.org>
Subject: NCBI check: contamination still present

Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).

Can you guys remove these and check the other genomes ASAP?

Thanks,

Matt

SUBID BioProject BioSample Organism ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []
We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.
After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondr
Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you s
Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations
Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NG

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: NCBI check: contamination still present
Date: June 18, 2017 at 6:02 PM
To: Ellison, Aaron aellison@fas.harvard.edu

Hey Aaron, they may need to trim the Ns and the end of the sequences too, I didn't check this, but the end of the error report lists the adapters tha

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

On Jun 18, 2017, at 5:43 PM, Ellison, Aaron <aellison@fas.harvard.edu> wrote:

> It looks like this is a consequence of too many "N"s. Is that the correct interpretation?
>
> Thanks for resolving this as quickly as possible,
> Aaron
>
> ---
>
> **From:** Lau, Matthew K.
> **Sent:** Sunday, June 18, 2017 17:33
> **To:** Jim Bochicchio <jboch@broadinstitute.org>; Ellison, Aaron <aellison@fas.harvard.edu>; Sarah Young <stowey@broadinstitute.org>; Caroline Cusick <ccusick@broadinstitute.org>
> **Subject:** NCBI check: contamination still present
>
> Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).
>
> Can you guys remove these and check the other genomes ASAP?
>
> Thanks,
>
> Matt
>
> SUBID BioProject BioSample Organism ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []
> We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.
> After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial se
> Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should
> Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to ma
> Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00
>
> Matt
>
> Postdoctoral Research Fellow
> Harvard Forest
> Harvard University
> 324 N Main St
> Petersham, MA 01366
>
> Office: (978) 756-6165
>
> Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.
>
> -- Miles Kington

From: **Ellison, Aaron** aellison@fas.harvard.edu
Subject: RE: NCBI check: contamination still present
Date: June 18, 2017 at 5:43 PM
To: Lau, Matthew K. matthewklau@fas.harvard.edu, Jim Bochicchio jboch@broadinstitute.org, Sarah Young stowey@broadinstitute.org, Caroline Cusick ccusick@broadinstitute.org

AE

It looks like this is a consequence of too many "N"s. Is that the correct interpretation?

Thanks for resolving this as quickly as possible,
Aaron

From: Lau, Matthew K.
Sent: Sunday, June 18, 2017 17:33
To: Jim Bochicchio <jboch@broadinstitute.org>; Ellison, Aaron <aellison@fas.harvard.edu>; Sarah Young <stowey@broadinstitute.org>; Caroline Cusick <ccusick@broadinstitute.org>
Subject: NCBI check: contamination still present

Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).

Can you guys remove these and check the other genomes ASAP?

Thanks,

Matt

SUBID BioProject BioSample Organism ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []
We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.
After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequ
Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should s
Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to mask
Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB0084

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington

From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: NCBI check: contamination still present
Date: June 18, 2017 at 5:33 PM
To: Jim Bochicchio jboch@broadinstitute.org, Aaron M Ellison aellison@fas.harvard.edu, Sarah Young stowey@broadinstitute.org, Caroline Cusick ccusick@broadinstitute.org

Hi Jim et al., looks like there's still contamination still present in at least one of the genomes. This is the NCBI error report for SM-AZXXO (see below).

Can you guys remove these and check the other genomes ASAP?

Thanks,

Matt

SUBID BioProject BioSample Organism ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster miamiana []

We ran your sequences through our Contamination Screen. The screen found contigs that need to be trimmed and/or excluded. Please adjust the sequences appropriately and then resubmit your sequences.

After you remove the contamination, trim any Ns at the ends of the sequence and remove any sequences that are shorter than 200 nt and not part of a multi-component scaffold. Note that hits in eukaryotic genomes to mitochondrial sequences can be ignored when specific criteria are met. Those criteria are explained below. Note that mismatches between the name of the adaptor/primer identified in the screen and the sequencing technology used to generate the sequencing data should not be used to discount the validity of the screen results as the adaptors/primers of many different sequencing platforms share sequence similarity. []

Some of the sequences hit primers or adaptors used in Illumina or 454 or other sequencing strategies or platforms. Primers at the end of a sequence should be removed. However, if primers are present within sequences then you should strongly consider splitting the sequences at the primers because the primer sequence could have been the region of overlap, causing a misassembly.

Screened 35,725 sequences, 265,000,276 bp. Note: 4,964 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 8 sequences with locations to mask/trim (8 split spans with locations to mask/trim)

Trim: Sequence name, length, span(s), apparent source scaffold00094 712997 15373..15411 adaptor:NGB00843.1 scaffold00121 612406 387957..387996 adaptor:NGB00843.1 scaffold00264 303333 247297..247333 adaptor:NGB00843.1 scaffold00282 286384 154047..154082 adaptor:NGB00843.1 scaffold00424 169256 91684..91733 adaptor:NGB00843.1 scaffold00635 94765 70145..70194 adaptor:NGB00843.1 scaffold02309 19718 14150..14191 adaptor:NGB00843.1 scaffold03495 10772 8825..8858 adaptor:NGB00843.1

Matt

Postdoctoral Research Fellow
Harvard Forest
Harvard University
324 N Main St
Petersham, MA 01366

Office: (978) 756-6165

Knowledge is knowing that a tomato is a fruit, wisdom is not putting it in a fruit salad.

-- Miles Kington