From: **Lau, Matthew K.** matthewklau@fas.harvard.edu
Subject: Re: Uploading Harvard Forest Ant Genomes to NCBI
Date: May 7, 2017 at 7:46 AM
To: Jim Bochicchio jboch@broadinstitute.org

Hi Jim, thanks for looking into that.

Also, I got this message back from the NCBI submission system. Looks like they're detecting primers in the sequence and won't take then as is. FYI this is for the filtered scaffolds file.

How do you normally handle this?

Cheers,

Matt

SUBID BioProject BioSample Organism ---------- ---------- ---------- ---------- SUB2631470 PRJNA385595 Aphaenogaster rudis [] 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 41,978 sequences, 280,952,533 bp. Note: 5,749 sequences with runs of Ns 10 bp or longer (or those longer that 20 MB) were split before screening. 31 sequences with locations to mask/trim (31 split spans with locations to mask/trim) Trim: Sequence name, length, span(s), apparent source scaffold00005 2225571 167999..168032

adaptor:NGB00843.1 scaffold00006 2018001 704097..704165
adaptor:NGB00749.1 scaffold00030 1266775 211712..211745
adaptor:NGB00843.1 scaffold00042 1100881 629517..629569
adaptor:NGB00843.1 scaffold00058 943114 46164..46214
adaptor:NGB00843.1 scaffold00088 791676 200390..200454
adaptor:NGB00756.1 scaffold00091 800321 588034..588136
adaptor:NGB00756.1 scaffold00107 735789 464178..464211
adaptor:NGB00843.1 scaffold00111 716142 416338..416363
adaptor:NGB00843.1 scaffold00129 638859 560569..560638
adaptor:NGB00749.1 scaffold00161 532482 203712..203732
adaptor:NGB00843.1 scaffold00172 537399 493052..493081
adaptor:NGB00843.1 scaffold00189 501372 386427..386476
adaptor:NGB00756.1 scaffold00217 416334 33832..33869
adaptor:NGB00843.1 scaffold00223 406083 332633..332734
adaptor:NGB00756.1 scaffold00234 386048 195823..195917
adaptor:NGB00756.1 scaffold00297 301462 301443..301462
adaptor:NGB00843.1 scaffold00314 268491 262087..262129
adaptor:NGB00843.1 scaffold00322 272223 247324..247390
adaptor:multiple scaffold00439 185104 158718..158737
adaptor:NGB00843.1 scaffold00500 161576 96827..96918
adaptor:NGB00756.1 scaffold00517 155854 130073..130137
adaptor:NGB00756.1 scaffold00686 101626 83337..83371
adaptor:NGB00843.1 scaffold00855 88061 19571..19639
adaptor:NGB00749.1 scaffold01106 68440 22221..22268
adaptor:NGB00843.1 scaffold01199 63326 12903..12943
adaptor:NGB00843.1 scaffold01857 41944 23768..23832
adaptor:NGB00756.1 scaffold02161 35894 1409..1475 adaptor:multiple scaffold03308 12204 1..49 adaptor:NGB00843.1 scaffold06723 2074 1766..1831 adaptor:multiple scaffold07340 1275 1240..1275 adaptor:NGB00725.1

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 May 5, 2017, at 3:43 PM, Jim Bochicchio <jboch@broadinstitute.org> wrote:

> HI Matt
>
> Let me check with our analysis folks to see if they have come across an error similar to this in the past when doing submissions.
>
> Thanks
> Jim
>
> On Thu, May 4, 2017 at 8:29 PM, Lau, Matthew K. <matthewklau@fas.harvard.edu> wrote:
>> Hi Jim, hope you're enjoying spring so far.
>>
>> I'm emailing because I'm in the process of uploading the genome sequences up to NCBI's database and I'm getting an error from my assembly (.agp) files. The error message is:
>>
>> seq_id=: CFastaReader: Input doesn't start with a defline or comment around line 0.One or more sequences either lack a Sequence ID or the ID contains invalid characters.Allowed characters in Sequence ID include letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks (*), and number signs (#).Each sequence must be uniquely identified by a valid Sequence ID

Sequence ID:
seq_id=: No sequences could be read in input

Sorry to bug you about this, but I'm hoping there's an easy answer you guys commonly encounter.

Best,

Matt

Postdoctoral Fellow
Harvard Forest
Harvard University
324 N Main St.
Petersham, MA 01366
978-756-6165

"Knowledge is knowing a tomato is a fruit. Wisdom is not putting it in a fruit salad." — Miles Kingston

--
Jim Bochicchio
Sr, Product Manager
Broad Technology Labs
Broad Institute
75 Ames Street
Cambridge MA 02142, USA

Phone: 617-714-8513
Website: https://www.broadinstitute.org/btl