A survey of popular 'omics' assembly tools

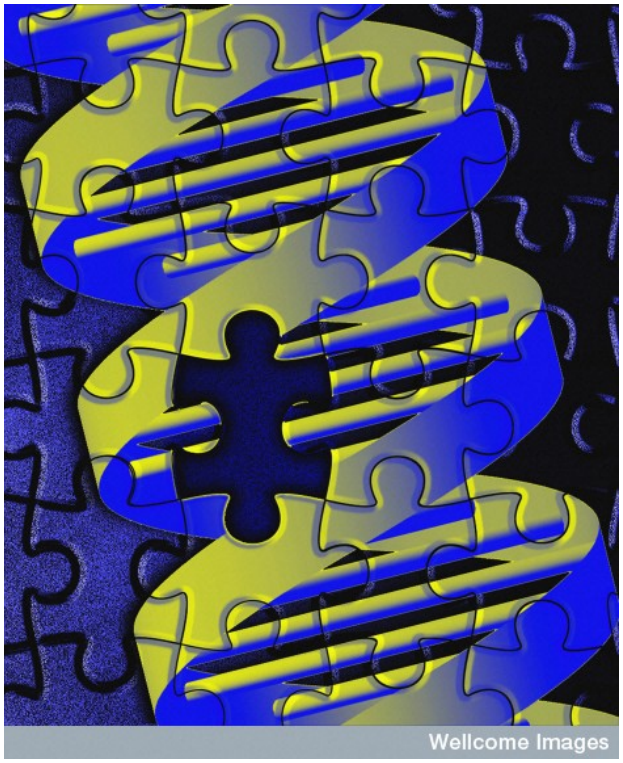


Image from Wellcome Images

Keith Bradnam

UC Davis Genome Center

July 2014

Overview

Between June 24th and July 1st 2014, I conducted a survey (using an online Google Form) to determine what tools are currently popular in the field of 'omics' assembly. The survey consisted of five questions, none of which were mandatory:

1. What is your preferred primary tool for eukaryotic genome assembly?
2. What is your preferred primary tool for bacterial/archaeal genome assembly?
3. What is your preferred primary tool for transcriptome assembly?
4. What is your preferred primary tool for metagenomics assembly?
5. How often do you generate genome/transcriptome/metagenome assemblies?

For the first four questions, I populated the survey form with a few examples of assembly tools but provided an option for 'Other'. During the survey period, I would regularly update the form to include choices that people had provided in the 'Other' category. For the last question, five choices were available:

- I haven't yet made an assembly
- Maybe 1–2 a year
- Maybe 1–2 a month
- Maybe 1–2 a week
- More than 1–2 a week

Although it is common to use more than one piece of software as part of a sequence assembly pipeline, the goal of this survey was to identify the *primary* assembly tools that are *currently* being used. This survey was promoted via my ACGT blog and via twitter.

Results

In the following tables, the names of assembly tools link to source web/FTP sites where the assembler software/code is available to download. For some commercial tools, the software is not directly available, but I try to link to the webpage that best describes the software.

Following the names of the software, I indicate a year that refers to the earliest date associated with the release of the software. This will sometimes correspond with the date of a publication that describes the software, but sometimes this date might precede the publication data if it is clear that the software was previously available (e.g. from developers website).

Other information is included below the tables as footnotes.

Table 1: Primary assembly tools for eukaryotic genome assembly

Assembler	Votes
ALLPATHS-LG (2010)	16
ABySS (2008)	13
SOAPdenovo (2009)	10
Velvet (2007)	10
SPAdes (2012)	7
Ray (2010)	6
Celera (2004)	5
CLC (2008)	5
MaSuRCA (2012)	5
Newbler (2008)[1]	5
Meraculous (2011)[2]	4
SGA (2010)	4
Other[3]	2
HGAP3 (2013)	2
SeqMan NGen (2007?)[4]	1
Minia (2012)	1
Platanus (2012)	1
Sprai (2013)	1

[1] There is not really an official Newbler website, so readers are advised to check out Lex Nederbragt's blog devoted to Newbler

[2] Note, this is a link to an FTP site (ftp://ftp.jgi-psf.org/pub/JGI_data/meraculous/)

[3] Two respondents indicated multiple tools (Celera/PBcR & ABySS/SOAPdenovo/Platanus/Newbler)

[4] This item was listed as 'DNASTAR' on the survey form (DNASTAR is the parent company).

Table 2: Primary assembly tools for bacterial/archaeal genome assembly

Assembler	Votes
SPAdes (2012)	29
Velvet (2007)	14
ABySS (2008)	6
ALLPATHS-LG (2010)	6
HGAP (2013)[1]	6
MIRA (1998?)[2]	6
SOAPdenovo (2009)	4
CLC (2008)	3
Meraculous (2011)[3]	3
SGA (2010)	3
IDBA-UD (2011)	2
A5 (2011)	1
Celera (2004)	1
MaSuRCA (2012)	1
Newbler (2008)[4]	1
Sprai (2013)	1

[1] Five people responded with 'HGAP2' and one person reported 'HGAP'

[2] Five people responded with 'MIRA' and one person reported 'MIRA 4'

[3] Note, this is a link to an FTP site (ftp://ftp.jgi-psf.org/pub/JGI_data/meraculous/)

[4] There is not really an official Newbler website, so readers are advised to check out Lex Nederbragt's blog devoted to Newbler

Table 3: Primary assembly tools for transcriptome genome assembly

Assembler	Votes
Trinity (2011)	57
SOAPdenovo-Trans (2011)	5
Trans-ABySS (2010)	4
CLC (2008)	3
MIRA	2
Oases (2010)	2
SeqMan NGen (2007?)[1]	1
IDBA (2010)[2]	1
IGC[3]	1
Pertran[3]	1
Rnnotator (2010)	1
Other[4]	1

[1] This item was listed as 'DNASTAR' on the original survey (this is the parent company).

[2] These were survey responses that may have intended to be submitted as IDBA-Tran (2012), a transcriptome assembler.

[3] I was unable to find a transcriptome assembler matching this name.

[4] One person responded with 'Velvet/Oases'

Table 4: Primary assembly tools for metagenome assembly

Assembler	Votes
Velvet (2007)[1]	17
Ray (2010)[2]	11
Ray-META (2012)[2]	6
SeqMan NGen (2007?)[3]	2
IDBA-UD (2011)[4]	2
META-IDBA (2011)[4]	2
Parallel-META (2011)	2
BAMBUS 2 (2011)	1
Celera (2004)	1
CLC (2008)	1
Genovo (2011)	1
Meta is[5]	1
MetaCortex (2013)	1

[1] It is possible that some of these respondents intended to specify MetaVelvet (designed for metagenomics) rather than just Velvet.

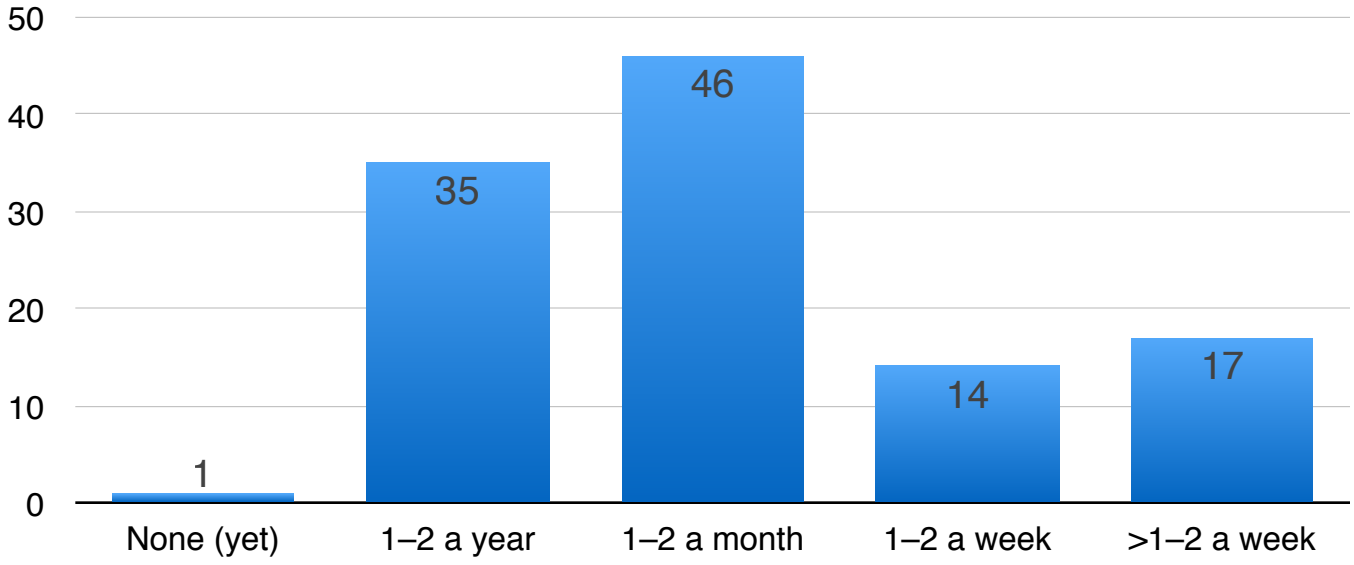
[2] It is possible that some respondents intended to specify Ray-META (designed for metagenomics) but instead chose Ray.

[3] This item was listed as 'DNASTAR' on the survey form (DNASTAR is the parent company).

[4] It is possible that some respondents intended to specify 'META-IDBA' (designed for metagenomics) but instead chose IDBA-UD.

[5] I was unable to find a transcriptome assembler matching this name.

Figure 1: Frequency of assemblies generated by survey respondents



Discussion

The results of this informal survey show that there is a great deal of variety as to which 'omics' assemblers are currently in vogue. Of all these 'omics' disciplines, it is transcriptome assembly where one tool, Trinity, most clearly outranks all others in terms of popularity. Metagenomics is currently dominated by two tools (Velvet and Ray/Ray-Meta), as is the case with bacterial/archaeal assembly (SPAdes and Velvet).

It is only the discipline of eukaryotic genome assembly where several tools compete for preeminence. The four most popular assembly tools in this category are relatively old (the newest tool is from 2010) compared to tools in the other categories.

It is important to conclude this survey by noting that **popularity may not equate to quality**. The sequence assembly tools which are currently the most popular may simply be those that:

1. Have received more attention in the scientific literature
2. Work with sequencing data from more platforms than other tools
3. Have more applications than other tools (e.g. genome *and* metagenome assembly)
4. Are easier to install
5. Can be run on the available computational resources (some assembly tools require large amounts of memory and/or many CPUs)