lociNGS: A Lightweight Alternative for Assessing Suitability of Next-Generation Loci for Evolutionary Analysis

Sarah M. Hird[1,2]*

[1] Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, United States of America, [2] Museum of Natural Science, Louisiana State University, Baton Rouge, Louisiana, United States of America

Abstract

Genomic enrichment methods and next-generation sequencing produce uneven coverage for the portions of the genome (the loci) they target; this information is essential for ascertaining the suitability of each locus for further analysis. lociNGS is a user-friendly accessory program that takes multi-FASTA formatted loci, next-generation sequence alignments and demographic data as input and collates, displays and outputs information about the data. Summary information includes the parameters coverage per locus, coverage per individual and number of polymorphic sites, among others. The program can output the raw sequences used to call loci from next-generation sequencing data. lociNGS also reformats subsets of loci in three commonly used formats for multi-locus phylogeographic and population genetics analyses – NEXUS, IMa2 and Migrate. lociNGS is available at https://github.com/SHird/lociNGS and is dependent on installation of MongoDB (freely available at http://www.mongodb.org/downloads). lociNGS is written in Python and is supported on MacOSX and Unix; it is distributed under a GNU General Public License.

Citation: Hird SM (2012) lociNGS: A Lightweight Alternative for Assessing Suitability of Next-Generation Loci for Evolutionary Analysis. PLoS ONE 7(10): e46847. doi:10.1371/journal.pone.0046847

Editor: Jason E. Stajich, University of California Riverside, United States of America

Received July 17, 2012; **Accepted** September 6, 2012; **Published** October 10, 2012

Copyright: © 2012 Hird. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by Google Inc. through the 2011 Google Summer Of CodeTM Program (http://code.google.com/soc/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The author was funded by Google Inc. to develop this open-source software. This does not alter the author's adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: shird1@tigers.lsu.edu

Introduction

To apply the immense sequencing capabilities of next-generation sequencing (NGS) technologies to population-level questions (i.e., those that require multi-locus, multi-individual data), genome enrichment methods are frequently employed. These methods aim to sample the genome at a reproducible subset of markers that can be obtained from many individuals and reduced to genotype (i.e., a set of phased alleles). Examples of these methods include amplicon sequencing [1], RAD-tags [2], complexity reduction of multilocus sequences (or CRoPS) [3] and sequence capture [4]; for a review of NGS methods suitable for multi-locus studies, see [5]. Genome enrichment methods often utilize a known or constructed reference for easing alignment of sequencing reads. Genotypes can then be called from the alignments, using a variety of bioinformatics methods (e.g., [6], [7]). This results in next-generation alignments to a reference and a set of loci for the individuals in the study; the loci can then be used in standard phylogeographic, phylogenetic or population genetic studies or other multi-locus analyses (e.g., [8];[9]). Prior to analysis, however, researchers must determine which loci are suitable for the questions being asked by assessing key parameters such as coverage and number of polymorphic sites or whether all populations are represented.

Current NGS file types are efficient at manipulating and storing alignment data but the parameters of interest are difficult to extract and can require custom bioinformatics scripts. Additionally, these file types are not useable in downstream analyses. Although large-scale, comprehensive programs like the Genome Analysis Toolkit (GATK) [10] can calculate coverage, if the parameters of interest are limited and include coverage per locus and coverage per individual, these programs are more heavy-duty and time-intensive than a user may want to invest. lociNGS is a lightweight, easy to use program that displays and outputs key parameters for researchers interested in multi-locus analysis of genotypes.

As more NGS papers come out, it should be standard to report summary statistics about coverage and polymorphism, in addition to the already standard number of total and high quality reads. Furthermore, as sequencing capacity continues to increase, the number of loci and number of individuals in a dataset will as well. Easily accessing, summarizing and reporting these parameters are important steps toward streamlining analysis and understanding large multi-locus datasets. lociNGS does not analyze any of the user-supplied data – it simply reports and exports summarized information about the dataset contained in the input files that is difficult to extract manually.

Methods

Overview

lociNGS was designed for use with multi-locus, multi-individual datasets generated through NGS. It collates information about loci, alignments and demographic data so that users can view summarized information about the genetic data (Table 1; Fig. 1)

Table 1. lociNGS parameters for the summary screen (Sum; Fig. 1a) and the individual screen (Ind; Fig. 1b).

Screen[a]	Parameter[b]	From[c]	*[d]	Definition
Sum	Individual	Demo		The individual's name
Sum	Population	Demo		The individual's population of origin
Sum	numLoci	Align	*	The number of loci called for each individual
Sum	totalReads	Align		Total number of reads sequenced in each individual
Sum	usedReads	Align		Total number of reads used for calling loci in this individual
Sum	percentUsed	lociNGS		UsedReads/TotalReads
Ind	LocusName	Loci		The name of the locus
Ind	Length	Loci		Number of bases in the locus
Ind	SNPs	Loci		Number of polymorphic sites
Ind	Number_Inds	Loci		Number of individuals called for this locus
Ind	Coverage_This_Ind	Align	**	Coverage for this locus in this individual
Ind	Coverage_Total	Align	**	Total coverage across individuals for this locus
Ind	Coverage_Used	Align	**	Total coverage for all individuals used in final locus

[a] Which screen the data are displayed on, the summary or the locus screen.
[b] Column header displayed in program; see Figure 1.
[c] Which input file the data are derived from, demographic data (demo), SAM/BAM alignments (Align), multi-FASTA locus files (loci) or calculated by lociNGS.
[d] * indicates this column's data serve as a button to pull up locus screen; **indicates this column's data serves as a button to print the corresponding reads to a multi-FASTA file.

doi:10.1371/journal.pone.0046847.t001

on the same screen as taxonomic and field data (e.g., subspecies, sampling locality, gender, etc.). In this way, one may assess the suitability of the data for further analysis.

The program has two types of display screens, both in table format. The "summary screen" contains demographic data, number of loci per individual (numLoci), total number of reads sequenced, number of reads used (along with the percentage of total). The numLoci data serve as buttons that open the corresponding "individual screen". This screen displays specific information about all the loci found in an individual, including length of the locus, number of polymorphic sites, number of individuals sequenced for that locus and coverage (for the individual, for all individuals, and for only the individuals with high enough coverage to be called). Each of the coverage categories serves as buttons that print the corresponding raw data in multi-FASTA format.

Program Input

lociNGS takes three categories of input: NGS alignment files, locus files (Fig. 2) and a demographic data file. When using genomic enrichment methods (or genome assembly methods), an alignment of the raw sequencing reads to a reference genome is often constructed using clustering or alignment programs (e.g., Geneious [11], Galaxy [12], Velvet [13], etc.). One common format for these alignments is SAM (Sequence Alignment/Map [14]) format or its binary version, BAM. These alignments contain a lot of information about the sequences and are lociNGS's source for many of the coverage and sequence data parameters (see Table 1). For input to the program, the alignment files need to be in sorted, indexed BAM format; the program samtools [14] can be used to convert SAM to BAM, sort and index the reads, if necessary.

Many traditional evolutionary analyses require individual loci that contain phased, homologous alleles for the individuals in the dataset. To get from alignments to loci, genotype-calling software is required (e.g. PRGmatic [6], STACKS [7], GATK [10,15], etc.). The loci are analogous to traditional Sanger sequencing loci and should be in multi-FASTA format. The locus files are the source for the SNP parameter as well as the locus names and length (see Table 1).

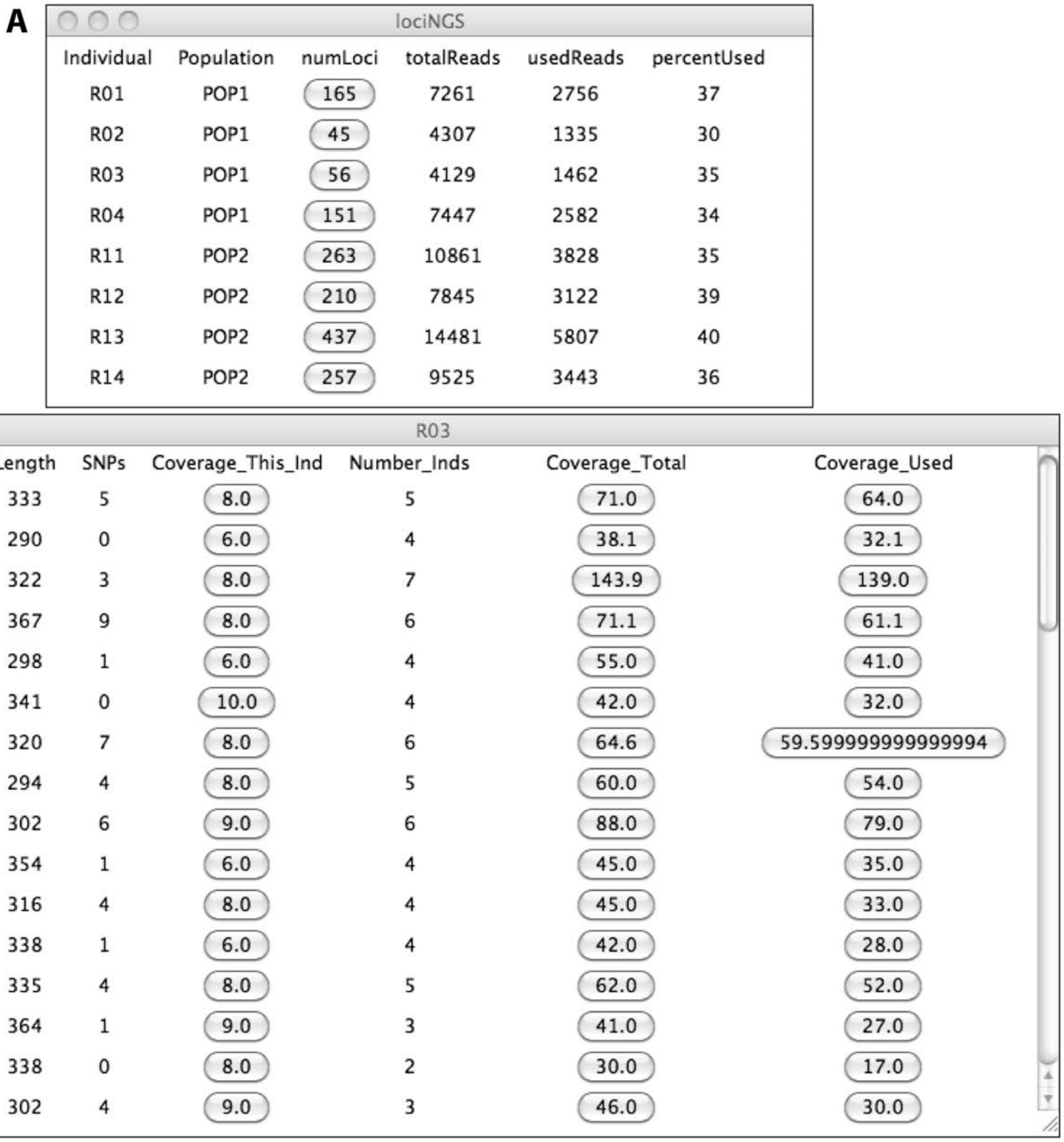
Finally, a demographic text file is required that, at a minimum, assigns each individual to a population; designating populations is frequently important in population level questions and is required because the output formats are capable of outputting a subset of populations or individuals. However, if this information is unknown or the user does not need the IMa2 or migrate output options, population can be set to something meaningless and the program will function properly.

Program Output

lociNGS outputs several different types of data. First, a table of all the information displayed to the user may be printed as a tab-delimited text file. This can then be edited with a spreadsheet or text-editing program to calculate averages, construct graphics, sort the data, etc.

Second, the raw sequences that were used to call a locus may be exported for an individual, for all individuals or just the individuals that were used in the final dataset; this information is contained in the alignment files but difficult to extract manually. These data are FASTA formatted.

Third, users may reformat a subset of populations or individuals into NEXUS [16], IMa2 [17] or Migrate [18] formats. These three formats are highly specific and are used in population genetics programs that can analyze large, multi-locus datasets. In addition, these formats can be rather time consuming to produce by hand or require custom scripts to produce for more than a few loci. lociNGS automates and combines the selection of loci and the construction of the appropriate input files. Under the export menu of the program, users select either populations or individuals they would like to include in the output of these formats; lociNGS then searches all the loci that contain at least one individual from the populations selected or all individuals selected.



Individual	Population	numLoci	totalReads	usedReads	percentUsed
R01	POP1	165	7261	2756	37
R02	POP1	45	4307	1335	30
R03	POP1	56	4129	1462	35
R04	POP1	151	7447	2582	34
R11	POP2	263	10861	3828	35
R12	POP2	210	7845	3122	39
R13	POP2	437	14481	5807	40
R14	POP2	257	9525	3443	36

Locus Name	Length	SNPs	Coverage_This_Ind	Number_Inds	Coverage_Total	Coverage_Used
maleyData_118	333	5	8.0	5	71.0	64.0
maleyData_124	290	0	6.0	4	38.1	32.1
maleyData_125	322	3	8.0	7	143.9	139.0
maleyData_134	367	9	8.0	6	71.1	61.1
maleyData_140	298	1	6.0	4	55.0	41.0
maleyData_141	341	0	10.0	4	42.0	32.0
maleyData_142	320	7	8.0	6	64.6	59.599999999999994
maleyData_143	294	4	8.0	5	60.0	54.0
maleyData_144	302	6	9.0	6	88.0	79.0
maleyData_145	354	1	6.0	4	45.0	35.0
maleyData_146	316	4	8.0	4	45.0	33.0
maleyData_151	338	1	6.0	4	42.0	28.0
maleyData_152	335	4	8.0	5	62.0	52.0
maleyData_154	364	1	9.0	3	41.0	27.0
maleyData_157	338	0	8.0	2	30.0	17.0
maleyData_159	302	4	9.0	3	46.0	30.0

Figure 1. Screen shots of lociNGS. Data include 8 individuals (rails); summarized data for the whole dataset shown in the summary screen (**A**) and one example of an individual (R03) screen shows parameters associated with individuals (**B**). Details of the column headings are in Table 1.
doi:10.1371/journal.pone.0046847.g001

The location of all exported files is logged to the screen and each has a unique file name.

Test Data

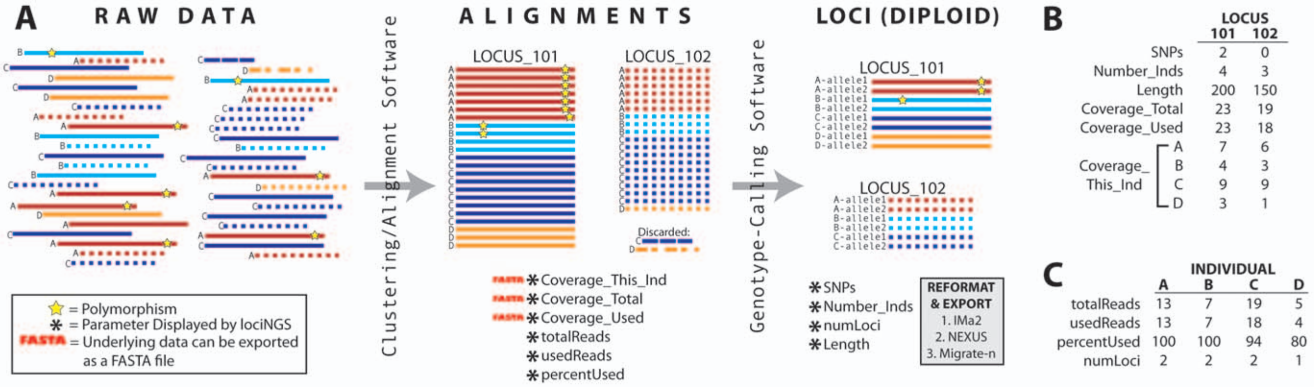
There is a small test dataset provided with the lociNGS distribution. This dataset includes four individuals at five loci. A copy of the exact parameter values displayed by lociNGS with the test data is included as supplemental material (Table S1).

Program Implementation

lociNGS is written in Python for a Unix-based system (e.g., MacOSX). It requires MongoDB as a separately installed program. lociNGS uses the Tkinter class of Python for a user-friendly graphical user interface. A modified version of seqlite (available: http://www.mbari.org/staff/haddock/scripts/) calls polymorphic sites from the aligned locus files; this tool works by simply counting the variable sites in an aligned FASTA file. The BAM files are not considered in the number of SNPs. The User Manual is included as a Supplementary File (Document S1).

An Example: Using lociNGS in Phylogeography

For many evolutionary analyses, a phased set of alleles is required as input; many NGS molecular and computational methods are now capable of producing such datasets. For example, McCormack et al. [8] generated restriction-digested fragments sequenced on a Roche 454 platform for two species of rails (*Rallus longirostris* and *R. elegans*) to identify fixed genetic differences in a bird hybrid zone; in this section I walk through a subset of their dataset that contains 4 individuals from each species (*R. longirostris* = R01, R02, R03, R04; *R. elegans* = R11, R12, R13, R14). The data was quality controlled and analyzed with PRGmatic [6], then loaded in to lociNGS. The summary screen



		LOCUS 101	LOCUS 102
SNPs		2	0
Number_Inds		4	3
Length		200	150
Coverage_Total		23	19
Coverage_Used		23	18
Coverage_ This_Ind	A	7	6
	B	4	3
	C	9	9
	D	3	1

	INDIVIDUAL A	B	C	D
totalReads	13	7	19	5
usedReads	13	7	18	4
percentUsed	100	100	94	80
numLoci	2	2	2	1

Figure 2. How the data are generated, where the parameters come from and example data. (**A**) Letters represent individuals and lines represent sequences; there are four individuals and two loci. Raw data from the sequencer is put through an alignment or clustering program to collect reads into alignments. From each alignment file, lociNGS reports totalReads, usedReads, percent reads used (percentUsed), Coverage_This_Ind, Coverage_Total and Coverage_Used; lociNGS will also export the data underlying the coverage parameters in FASTA format. Genotype-calling software will reduce sequence reads to loci (phased alleles). lociNGS uses these loci to report SNPs, Number_Inds, numLoci and Length; the program can reformat the loci into IMa2, NEXUS or Migrate formats. For further explanation of the parameters, see Table 1. (**B**) The parameter values for the two loci (LOCUS_101 and LOCUS_102) in this example. (**C**) The parameter values for the four individuals (A,B,C,D) in this example.
doi:10.1371/journal.pone.0046847.g002

(Fig. 1a), which can be exported as a tab-delimited text file, informs the user of how efficient the method was, in terms of how many reads were aligned to the reference genome compared to total number of reads (Fig. 3). It also displays the total number of loci that each individual belongs to; these data functions as a button that opens the individual screen for the given individual (Fig. 1b).

The individual screen contains detailed information about each of the loci with links to the raw data that make up each locus (Fig. 1b). Exporting this data as a tab-delimited text file allows the user to determine the distributions of polymorphic sites (Fig. 4a), number of individuals (Fig. 4b) and coverage per individual (Fig. 4c) across all loci. One can also assess how well each individual performed, by calculating average coverage. One may use this information to decide which individuals are worth resequencing with custom primers (to fill in their data matrix) or how to prune their dataset to the most complete or informative loci.

If a particular locus has more polymorphic sites than one might expect by the processes of natural selection or drift, the user can output the sequence reads that compose the raw data to investigate underlying copy number. With the raw read data, an alignment and phylogenetic tree can be estimated from either a single locus for one individual or all the reads underlying a single locus from all individuals (Fig. 5), but analysis of the raw reads is up to the user. For these data, I used Muscle [19] for alignment (using all defaults) and Geneious [11] to construct a neighbor-joining tree (using an HKY model of genetic distance and no outgroup). An analysis like this is very quick and although more sophisticated phylogenetic algorithms exist, for the purposes of assessing number of clades, these methods worked well. Once a tree has been constructed, if there are two (or fewer) major clades for each individual, it is likely that the sequences derive from a single diploid locus (Fig. 5a). However, if there are more clades than the ploidy of the organism allows, there may be multiple genomic sources of the data (Fig. 5b). One can also assess paralogy in the reads from all individuals at a locus: if all the reads from each individual belong to two or fewer clades, the locus is likely single copy (Fig. 5c). However, if one or more individuals belong to multiple clades, the underlying copy number may not be one (Fig. 5d).

Finally, lociNGS will export the data in three formats for input to evolutionary analysis programs. Users select exportation of either individuals or populations. The program searches for all loci that contain at least one individual from each of the selected categories. In other words – if all individuals are selected, only the loci that contain all individuals will be reformatted and printed. If

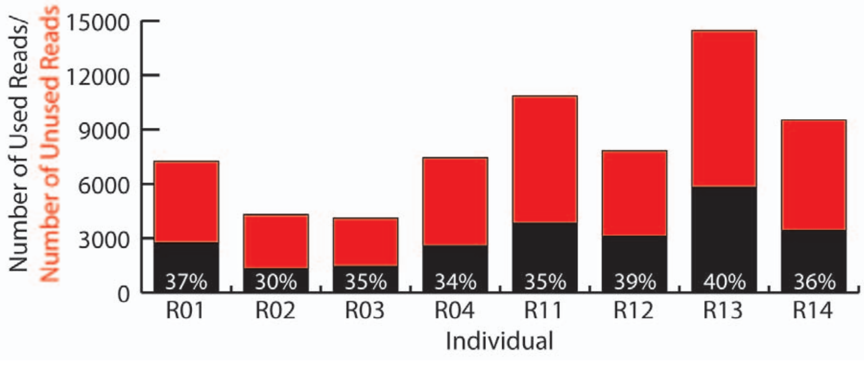


Figure 3. Number of reads per individual. Black portion of bars represents reads aligned to the reference; red portion accounts for unused reads. Percentage of reads used is shown in white text.
doi:10.1371/journal.pone.0046847.g003

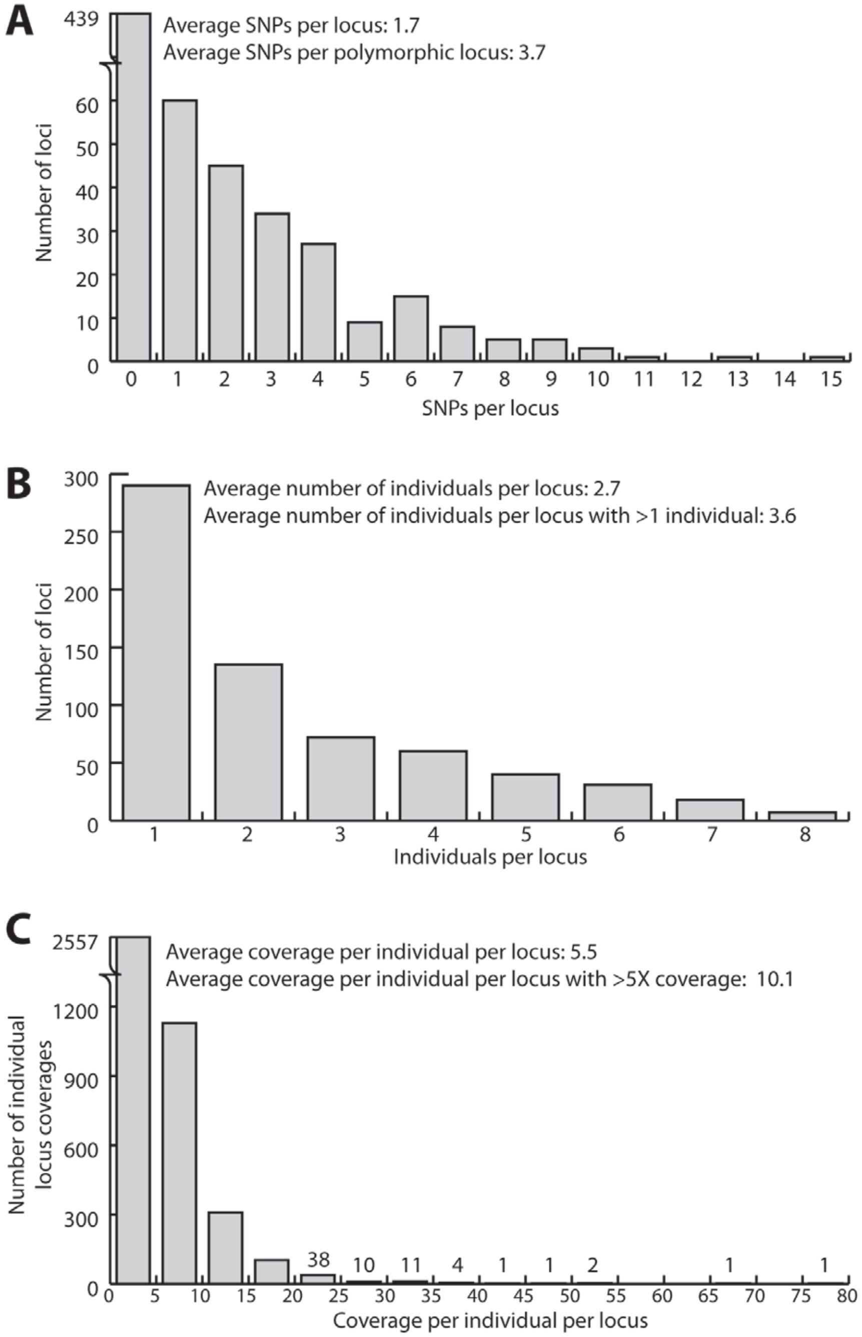


Figure 4. Summary histograms of important parameters in the rail dataset. Number of polymorphic sites (**A**), individuals present in each locus (**B**), individual coverage on a per locus basis (**C**). Note the scale of the dependent axis changes on (**A**) and (**C**).
doi:10.1371/journal.pone.0046847.g004

all populations are selected, only the loci that contain at least one individual from each population will be reformatted and printed.

Altogether, these simple functions provide the user with an overall sense of how their method and data perform at a basic level.

Conclusions

With the ever-increasing amount of data that is gathered with NGS, it is important to assess the suitability of the reads for further analysis. lociNGS provides a simple and quick way to determine which loci and which individuals have enough coverage and polymorphism to use in evolutionary analysis. Furthermore, the program automatically converts suitable loci to several file formats that are common in evolutionary analysis and time consuming when done by hand. Small, easy to use programs designed for a specific task allow researchers to customize their workflow and minimize or eliminate the learning curve for complex programs.

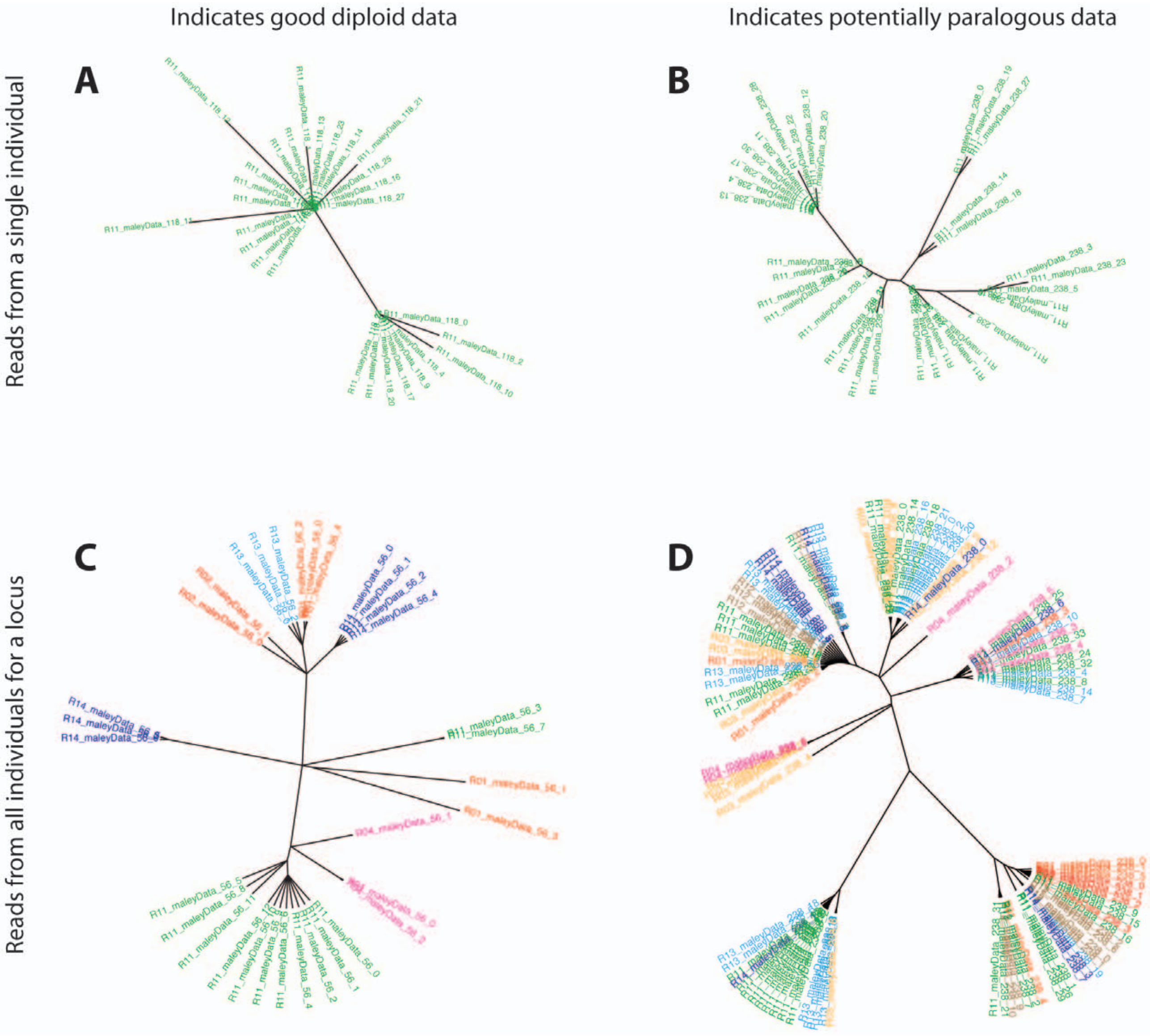


Figure 5. Neighbor-joining trees of aligned reads (reads output from the program) to help assess copy number. Shown are reads from one individual (**A**, **B**) and all the reads for a locus (**C**, **D**). Both (**A**) and (**C**) imply single copy loci; in (**A**) there are only two major clades and in (**C**) the reads for each individual, as shown by the different colors, belong to two clades at the most. Both (**B**) and (**D**) indicate potential multi-copy loci; in (B), there are greater than two clades and in (**D**) the reads for each individual, as shown by the different colors, are frequently distributed across greater than two clades.
doi:10.1371/journal.pone.0046847.g005

Supporting Information

Table S1 Expected results from test data included with lociNGS. The exact results that the program should output if the test data is input into the program.
(PDF)

Document S1 README file for lociNGS. The README file contains detailed information about installation, input options, output options and troubleshooting.
(TXT)

Acknowledgments

My appreciation goes to JM Brown and BA Chapman for mentorship and coding help on this project; additional appreciation to Google Summer of CodeTM and National Evolutionary Synthesis Center (NESCent) for sponsoring this project and offering guidance, particularly H Lapp, KA Cranston, T Vision and C Smith. JM Maley was a helpful beta tester. Thank you to TA Pelletier, JM Brown, J Stajich and four anonymous reviewers for insightful comments on various versions of this manuscript; JM Maley, JE McCormack and RT Brumfield graciously supplied the rail dataset.

Author Contributions

Conceived and designed the experiments: SMH. Performed the experiments: SMH. Analyzed the data: SMH. Contributed reagents/materials/analysis tools: SMH. Wrote the paper: SMH.

References

1. Binladen J, Gilber M, Bolback J, Panitz F, Bendixen C, et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS One 2: e197.
2. Baird N, Etter P, Atwood T, Currey M, Shiver A, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: 3376.
3. van Orsouw N, Hogers R, Janssen A, Yalcin F, Snoeijers S, et al. (2007) Complexity Reduction of Polymorphic Sequences (CRoPS): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS One 2: 1172.
4. Okou D, Steinberg K, Middle C, Cutler D, Albert T, et al. (2007) Microarray-based genomic selection for high-throughput resequencing. Nature Methods 4: 907–909.
5. McCormack J, Hird S, Zellmer AJ, Carstens B, Brumfield R (2012) Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution In Press.
6. Hird S, Brumfield R, Carstens B (2011) PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. Molecular Ecology Resources 11: 743–748.
7. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1: 171–182.
8. McCormack J, Maley JM, Hird S, Derryberry E, Graves G, et al. (2012) Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. Molecular Phylogenetics and Evolution 62: 397–406.
9. Zellmer AJ, Koopman MM, Hird SM, Carstens BC (2012) Deep Phylogeographic Structure and Environmental Differentiation in the Carnivorous Plant *Sarracenia alata*. Systematic Biology, 61(5):763–777.
10. Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20: 1297–1303.
11. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, et al. (2012) Geneious v5.6. Available: http://www.geneious.com. Accessed 2012 Jun 22.
12. Goecks J, Nekrutenko A, Taylor J, Team TG (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology 11.
13. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.
15. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43: 491–498.
16. Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: an extensible file format for systematic information. Systematic Biology 46: 590–621.
17. Hey J (2010) Isolation with migration models for more than two populations. Molecular Biology and Evolution 27: 905.
18. Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152: 763–773.
19. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 6.