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 NGSS methods suitable for multi-locus studies, see [5].

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. loqiNGS 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. loqiNGS 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

loqiNGS 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).
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.

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

| Screen | Parameter | From | *a* | 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.

*a*Which input file the data are derived from, demographic data (demo), SAM/BAM alignments (Align), multi-FASTA locus files (loci) or calculated by LOCiNGS.

** 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.

![Image of Table 1](https://example.com/table1.png)

[1] 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.
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 LOCI NGS distribution. This dataset includes four individuals at five loci. A copy of the exact parameter values displayed by LOCI NGS with the test data is included as supplemental material (Table S1).

Program Implementation
LOCI NGS is written in Python for a Unix-based system (e.g., MacOSX). It requires MongoDB as a separately installed program. LOCI NGS uses the Tkinter class of Python for a user-friendly graphical user interface. A modified version of seqsite (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 LOCI NGS 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 into LOCI NGS. The summary screen

Figure 1. Screen shots of LOCI NGS. 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
lociNGS: Summarizing Next-Gen Loci

(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 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](https://journals.plos.org/plosone/article/file?jid=pone&vid=10.1371%2Fjournal.pone.0046847&did=6.1371%2Fjournal.pone.0046847&pdf=true)

![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](https://journals.plos.org/plosone/article/file?jid=pone&vid=10.1371%2Fjournal.pone.0046847&did=6.1371%2Fjournal.pone.0046847&pdf=true)
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.
**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.

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

**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 Osouw 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 ‘provi-
sional-reference genome’. Molecular Ecology Resources 11: 743–748.
7. Catchen JM, Amores A, Hoehnlohe P, Uresko 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: 395–406.
9. Zellmer AJ, Koopenan MM, Hird SM, Carstens BC (2012) Deep Phylogeo-
graphic 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)
Genetic v3.6. Available: http://www.genetics.com. Accessed 2012 Jun 22.
12. Goecks J, Nekrutenko A, Taylor TG (2010) Galaxy: a comprehensive
approach for supporting accessible, reproducible, and transparent computa-
tional 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: 829–839.
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: 491–496.
16. Maddison DR, Swoford 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.