WHITEPAPER

Dovetail *De Novo* Genome Assembly Service Process

Dovetail GENOMICS

Table of Contents

Overview Of The Dovetail *De Novo* Genome Assembly Process

Dovetail Genomics™ offers the solution to generating well-scaffolded and accurate genome assemblies for any organism. Genomes are assembled in most cases using data from only two types of short-read, paired-end sequencing libraries: conventional shotgun libraries and Dovetail's proprietary Chicago™ libraries. Existing assemblies can also be improved with the simple addition of a Chicago library and scaffolding with our HiRise™ software pipeline. The process is simple, effective, flexible, and affordable. This white paper describes the method, from assay to analysis, as well as supporting data and case studies.

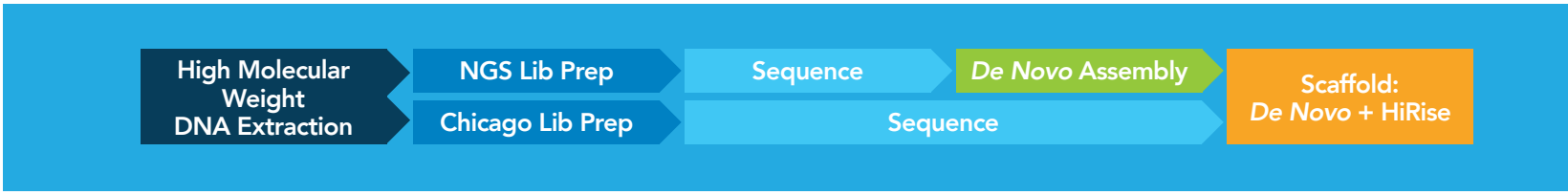
The key to Dovetail's genome assembly process is the information provided by our Chicago libraries. Uniquely, these libraries yield read pairs with separations up to the maximum fragment size of the input DNA. Libraries can be prepared with as little as 500 ng of input DNA of the highest possible molecular weight. Chicago libraries are prepared by Dovetail as a service for our customers from virtually any source of DNA. Input DNA for the library preparation may be provided by you directly, or Dovetail's scientists can use their extensive experience to extract high-quality, high-molecular-weight (HMW) DNA from your biological samples. Once prepared and QC'd, Chicago libraries can be sequenced by Dovetail or by a provider of your choice. Typical sequencing depth requirements range between 150 and 300 million read pairs (1-2 HiSeq 2500 lanes in Rapid Run Mode).

Our service is similarly flexible with regard to the starting or draft input genome assembly. Virtually any draft assembly can be used as input to the HiRise software pipeline. Alternatively, if you have no draft assembly or would simply like to start from scratch, Dovetail will produce a draft *de novo* assembly for you. This can be done from data that you provide or from data collected on your behalf by Dovetail.

With the draft assembly and Chicago data in hand, the final phase is scaffolding with our in-house HiRise software pipeline. This pipeline has been purpose-built to leverage the unique information provided by Chicago libraries for scaffolding. Our team of experienced bioinformaticians will perform the scaffolding process for you, tuning as necessary to yield the best possible results.

Upon completion of the service you will receive your new assembly along with a report summarizing the improvement over the draft or *de novo* assembly and key assembly statistics. Additionally, all raw data generated for the project will be provided as well as detailed maps showing the relationships between contigs and scaffolds from the input assembly to those in the HiRise assembly. Finally, breaks made to the input assembly on the basis of the Chicago library information are also reported in detail.

Figure 1: An overview of Dovetail's assembly service process.



The Dovetail Chicago Method

The core of Dovetail's assembly service offerings is the long-range genomic information yielded by Dovetail's proprietary Chicago libraries. An overview of the library generation process is diagrammed in Figure 2 and described here.

The assay begins with high molecular weight (50+ kbp) input DNA, depicted in Panel A as a bold black line. The first step is to reconstitute chromatin from the input DNA using in vitro chromatin assembly. This step utilizes purified nuclear remodeling and chaperone proteins and histones to convert the input naked DNA into chromatin (DNA which is wound around and tightly associated with histones). This step is depicted in Panel B where the now-complexed histones are indicated by blue circles upon the DNA. Re-association of the DNA with proteins in a uniform manner lays the foundation for later crosslinking steps. Chromatin reconstitution works with any source of DNA, even those with no native chromatin (e.g., bacteria).

The next stage is crosslinking (fixation) of the reconstituted DNA, as depicted in Panel C. Addition of a fixative agent (e.g., formaldehyde) produces crosslinks (covalent connections, depicted as red lines between histones) among the histones associated with the DNA. Crosslinking condenses the previously long, linear strand of chromatin into a globular chromatin aggregate and stabilizes it. The crosslinking serves two critical functions in the library preparation process. The first is bringing linearly distant portions of the molecule (e.g., the ends of the DNA) into close spatial proximity. The second is stabilization. Subsequent steps in the process will cut the DNA and the crosslinks retain the association of the now many DNA fragments originating from a single large initial fragment.

As depicted in Panel D, the crosslinked chromatin is then digested with a restriction endonuclease to generate sticky-ended fragments. This cutting step generates many new ends to participate in the subsequent ligation step. Critically, all of the fragments generated at this step originate from the same original large fragment and their association is maintained by the crosslinks.

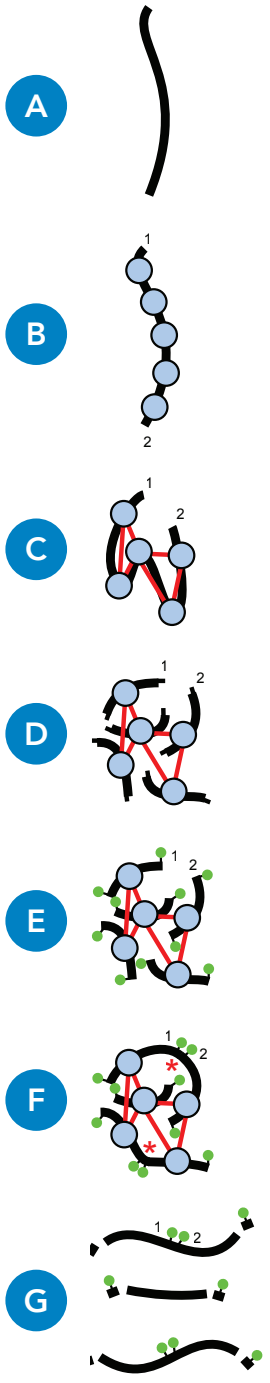
To prepare for ligation, the newly-generated sticky ends are made blunt with a polymerase fill-in that includes a biotinylated nucleotide (Panel E, biotins are green circles). Biotins mark the ends for a later enrichment step.

Next, a DNA ligase is added to perform blunt-end ligation of the many ends within a given chromatin aggregate (Panel F, ligations indicated with red asterisks). Many of these ligation events will result in the joining of two pieces of DNA that were not near to one another in the original fragment, but which have been brought into close spatial proximity by previous steps. For example, segments 1 and 2 were very distant in the original fragment (Panel B) but have become linked in F after condensation and ligation. It is the linking of these previously-distant fragments that captures the long-range information contained in the original molecule.

After ligation, chromatin is removed and DNA is purified and processed to remove biotin that is not internal to the resulting fragments, i.e. those which have not participated in a ligation event (Panel G). Finally, the resulting library is enriched for biotin-containing fragments through a streptavidin bead pull-down and a sequencing library is prepared.

The resulting libraries contain many "chimeric" fragments, i.e., single fragments composed of at least two originally distant fragments. The libraries can then be sequenced with full-length or paired-end sequencing and the two or more original fragments recovered and mapped. The genomic separation of the original fragments follows a well-modeled insert distribution, which provides the information used for scaffolding, analysis of structural variation, and phasing.

Figure 2:
An overview of the Chicago library preparation method.



Building A Genome Assembly With Dovetail's HiRise Scaffolder

HiRise is a sophisticated software pipeline developed specifically for scaffolding draft genome assemblies with sequence data from Chicago libraries. It is built around a statistical model that takes into account the unique features and read pair separations of these libraries. HiRise can dramatically improve assemblies constructed from virtually any assembler.

Shotgun assembly

For customers without a preexisting draft genome sequence for their project, Dovetail offers a *de novo* genome sequencing and assembly service. Our assembly pipeline is based on a customized version of the Meraculous assembler from the Department of Energy Joint Genome Institute (http://jgi.doe.gov/data-and-tools/meraculous). For most projects, Illumina sequencing of a single paired-end (PE) library is sufficient to create a starting assembly for HiRise. For some highly repetitive genomes it is necessary to generate mate-pair data to get an assembly with long enough scaffolds for scaffolding with HiRise.

Scaffolding With HiRise

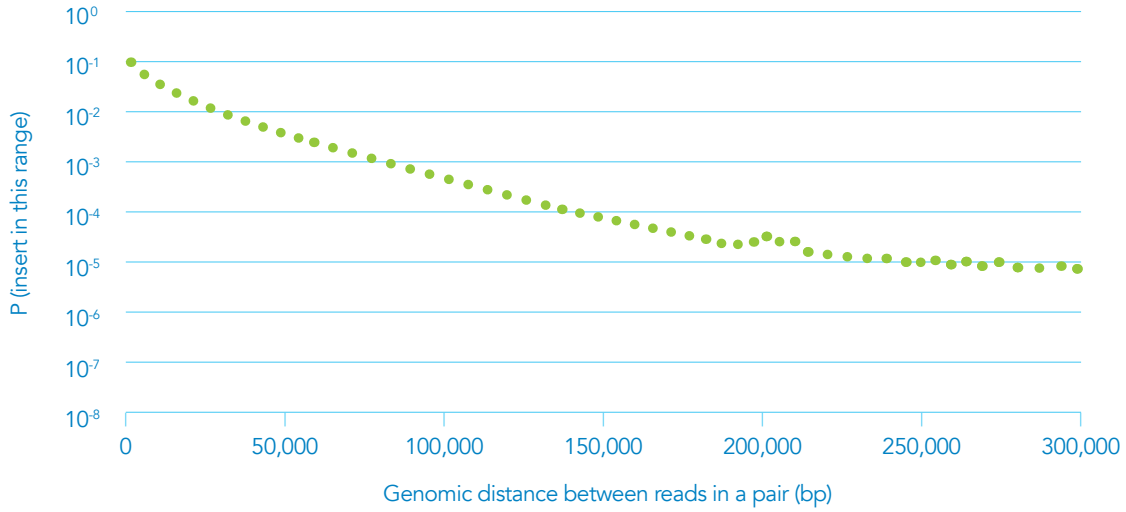
Overview

The HiRise assembler takes three inputs: 1) a draft genome assembly in FASTA format, 2) short insert shotgun read sequences in FASTQ format, and 3) Dovetail Chicago library sequencing reads in FASTQ format. The pipeline improves assemblies in three ways: 1) it orders and orients contigs and contig fragments into scaffolds, 2) it uses the pattern of mapped Chicago library reads to identify and break misjoins in the input assembly, and 3) it fills sequence gaps between contigs after scaffolding when the shotgun reads imply a unique sequence to close the gap.

Read mapping and likelihood model estimation

Dovetail uses the SNAP (http://snap.cs.berkeley.edu) read mapper to align both shotgun and Chicago reads to the input assembly. The frequency distribution of mapped Chicago read pair separations is used to calibrate the likelihood model on which the scores used in HiRise are based. An exemplary insert distribution for a human (NA12878) Chicago library is shown below to demonstrate the libraries' unique read pair separation distribution.

Figure 3: A typical read pair separation distribution from a human Chicago library.



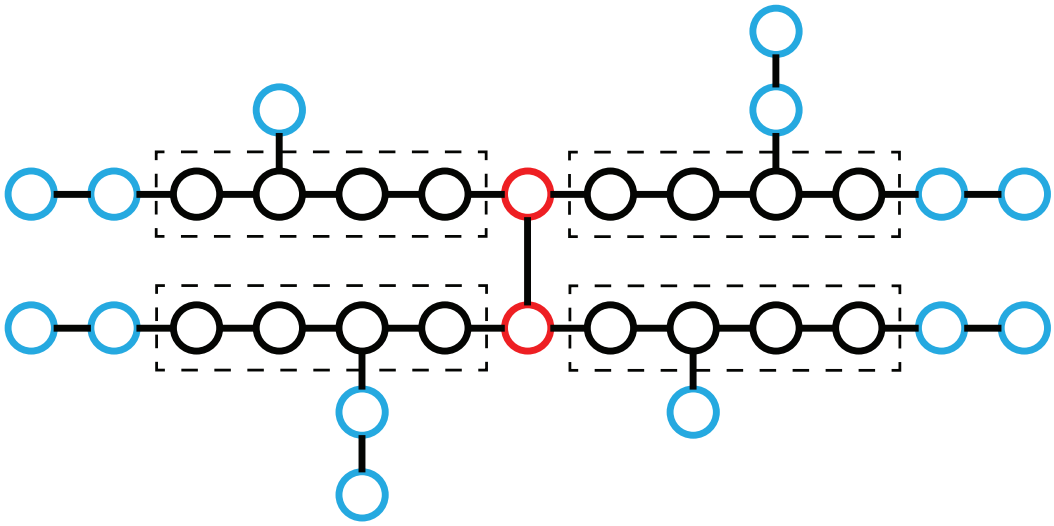
Sequence masking

HiRise uses two methods to identify repetitive segments of the input assembly: 1) excess depth of mapped shotgun reads, and 2) excess density of mapped Chicago reads. Chicago links in these repetitive segments are ignored because they are confounding to scaffolding.

Construction of seed scaffolds and iterative scaffolding

HiRise links fragments from the starting assembly into scaffolds iteratively, starting from a set of seed scaffolds. Seed scaffolds are constructed using a topological pruning algorithm (Figure 4) that 1) finds a minimum spanning tree *T* of a graph in which nodes are input fragments and an edge (a,b) is labeled with the Chicago read pairs that link fragments a and b, 2) constructs a processed graph *T'* by several rounds of removal of nodes with degree one, and 3) extracts as seed scaffolds the un-branched linear subgraphs of *T'*. In each round of iteration, likelihood ratio scores are computed for candidate scaffold merges (both end-to-end and by intercalation). Scaffold merges are implemented in decreasing order of log likelihood ratio (LLR) score magnitude down to a minimum score, as long as scaffold linearity constraints are not violated.

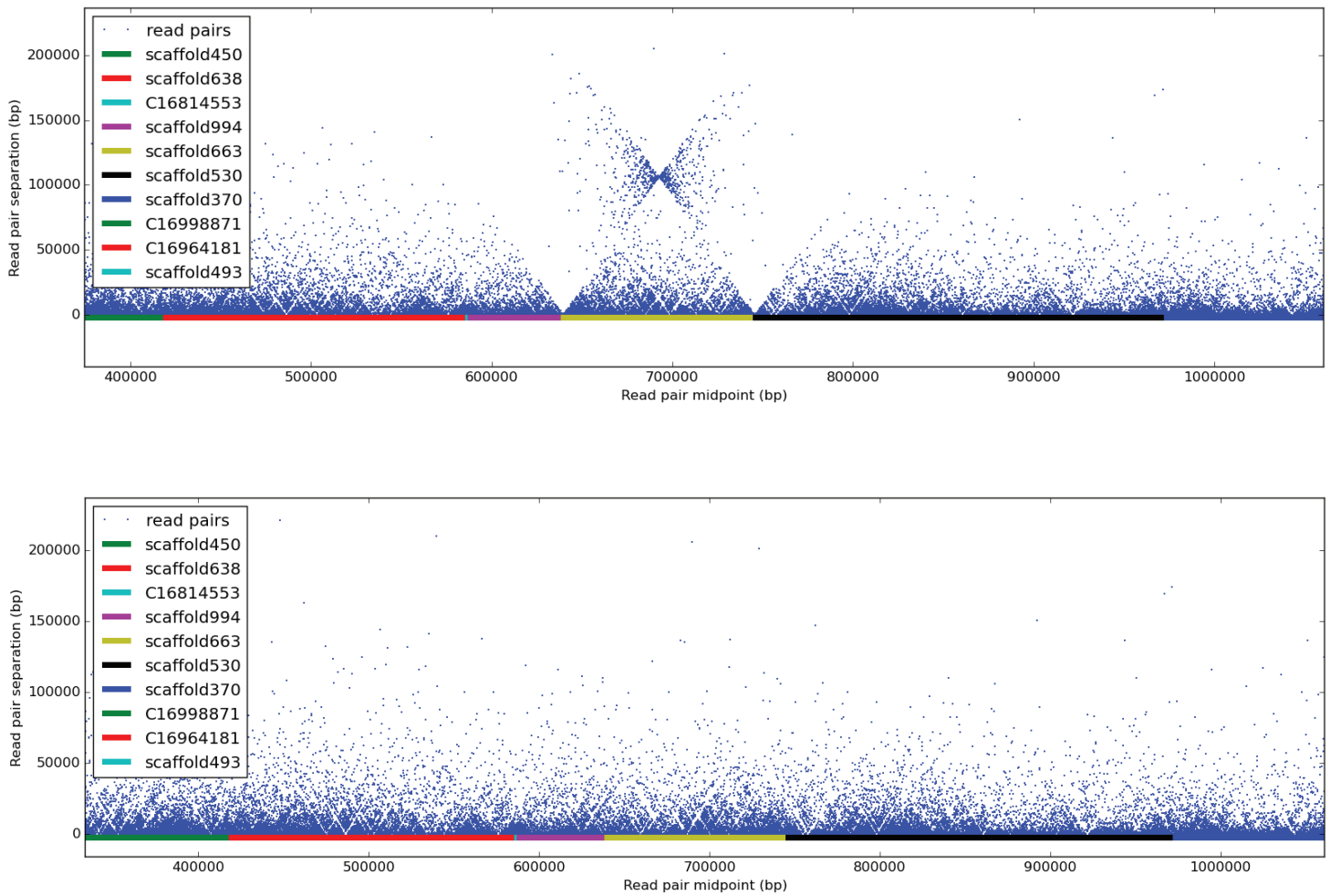
Figure 4: A diagrammatic representation of the initial stages of the scaffolding pipeline. In the diagram, nodes (circles) represent input assembly fragments (contigs/scaffolds). Connections between nodes represent Chicago pairs linking those nodes. The initial tree (*T*) includes all depicted nodes. The tree after pruning (*T'*) includes black and red nodes. Blue nodes (those connected to only one other node) are removed in pruning. Finally, nodes with more than two connections (red nodes) are removed, leaving the unbranched linear subgraphs surrounded by dotted lines.



Local refinement of contig order and orientation

Figure 5 (top) shows an artificially constructed example in which one fragment of the starting assembly (scaffold663) has been artificially inverted (for illustration purposes) within one of the HiRise scaffolds, giving rise to an anomalous pattern in the distribution of mapped chicago read pairs. As part of each scaffolding iteration, HiRise systematically computes an approximation to the total LLR score change that would result from every possible change of fragment ordering and orientation that does not move any fragment more than two steps from its starting rank in the ordering. The bottom panel shows the mapping pattern after order and orientation refinement.

Figure 5: Illustration of an inversion corrected by HiRise in order and orientation refinement.



Misjoin detection and breaking

In each round of iterative improvement, HiRise evaluates every position in each scaffold to identify assembly misjoins. Figure 6 shows the distribution of Chicago read pairs in the region of a misjoin identified in a customer-provided assembly scaffold. The absence of Chicago pairs linking across the misjoin provides a clear signature. Figure 7 overlays the LLR support score, which drops below zero at the site of the misjoin, indicating the higher likelihood under the model that the two sides of the scaffold are unlinked.

Figure 6: Figure demonstrating the lack of Chicago read pairs spanning a misjoin in the input assembly.

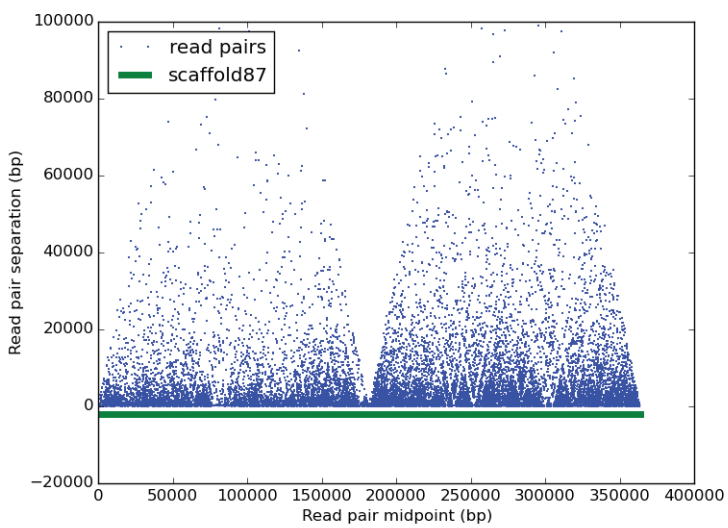
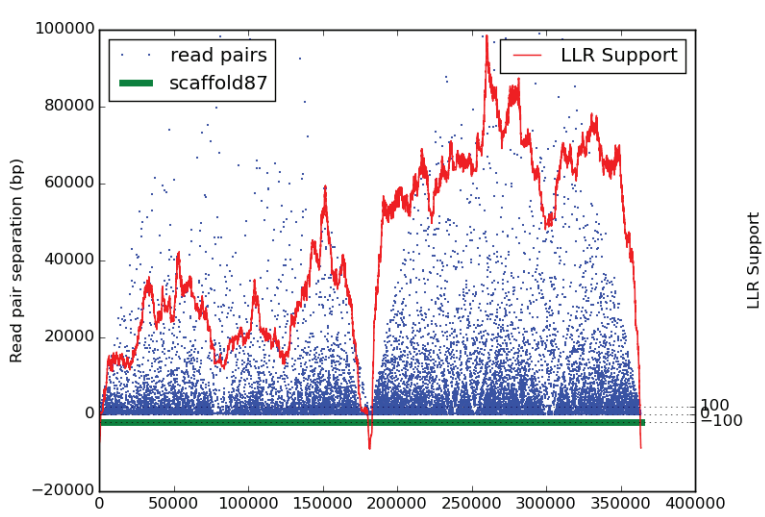


Figure 7: Same region as Figure 6 with LLR score overlaid.



HiRise outputs

When your improved genome assembly is returned to you, you will receive: a FASTA file of the HiRise scaffold sequences, a table specifying how the starting contigs were broken and re-joined, a BAM file containing the mapping of Chicago library reads to the starting assembly, and a users' guide to interpreting these files.

Prairie Chicken: A Case Study

The heath hen was a charismatic bird endemic to the grasslands of New England and surrounding regions. It was driven to extinction in large part due to over-hunting by humans. As a key mediator of biodiversity in its habitat, the "de-extinction" of the heath hen offers promise for restoring balance to that environment. Consequently, the Revive & Restore foundation has set out to understand the challenges and feasibility of reviving and reintroducing the heath hen, potentially creating the first community-backed de-extinction effort.

A critical component of the heath hen revival is the reconstruction of its original genome. Because no living heath hens remain, the genomic "scaffold" of this reconstruction effort had to be constructed from a close, living evolutionary relative. The greater prairie chicken was selected as an ideal candidate due its close evolutionary relationship to the heath hen. At the outset of this project only limited genomic sequence information had been collected for the greater prairie chicken, and no full genome assembly existed. Consequently, Revive & Restore partnered with Dovetail to produce a high quality greater prairie chicken genome assembly to lay the foundation for the heath hen reconstruction efforts.

To begin, members of the heath hen project provided Dovetail with tissue samples from two female greater prairie chickens, shipped frozen immediately after sampling. Dovetail extracted DNA from a number of the samples and selected the best extraction (that with highest yield and largest fragment size) for the full assembly effort. DNA from the chosen sample was used to prepare both conventional shotgun libraries and Chicago libraries, the only two data types used in this project. The shotgun libraries were sequenced to a Q20 depth of ~95X, which corresponds to approximately 300 million read pairs at 2x150 bp. Dovetail then performed a Meraculous assembly with the collected shotgun data to yield a draft assembly with a scaffold N50 of 136 kbp.

Next, Dovetail produced and sequenced a Chicago library from the same source high molecular weight DNA. This library was sequenced to a physical coverage level of ~110X, corresponding to approximately 150 million read pairs at 2x100 bp. This library and the draft assembly were scaffolded by Dovetail's HiRise pipeline to yield a final scaffold N50 of 12.2 Mbp, a nearly 90-fold increase in contiguity. The final genome totaled nearly 1 Gbp in size.

The entire duration of this process, from sample reception to genome delivery, took only 8 weeks. The greater prairie chicken's genome is now in the hands of Revive & Restore's scientists, who are hard at work leveraging it to reconstruct the heath hen's genome.

	Draft (*De Novo*) Assembly	Final HiRise Assembly
N50	136 kbp	12.2 Mbp
Data	300 million 2x150 shotgun read pairs	+ 150 million 2x100 Chicago read pairs

Figure 8:
Greater prairie chicken

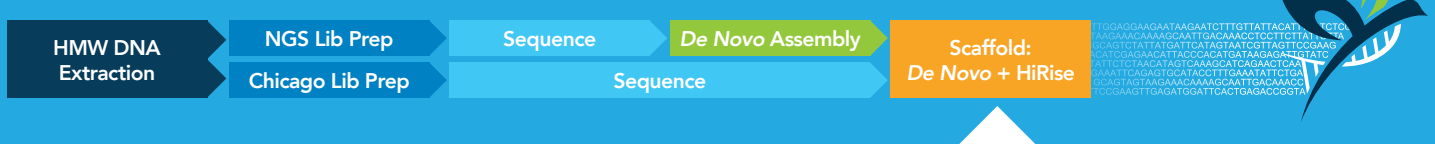
Figure 9:
Heath hen

Organismal Flexibility: Examples Of Genome Assemblies Built Using Dovetail's Process

Chicago libraries and the HiRise pipeline are capable of accommodating a wide variety of organisms. Dovetail has produced assemblies for plants, vertebrates, insects, and more. Below is a table of results from our beta program demonstrating significant improvements across a broad swath of life.

Table 1: A list of assemblies Dovetail has improved with its technologies.

Organism	Genome Size (Mbp)	Fold Physical Coverage (in 1-50 kbp bins)	Input N50 (kbp)	Final N50 (kbp)	Fold N50 Improvement
Vampire Bat	2,088	82x	5,498	13,814	3x
Cichlid	845	118x	1,208	3,395	3x
Potato	729	208x	755	5,868	8x
Butterfly	322	55x	143	3,707	26x
Pigeon	1,086	38x	70	3,739	54x
Prairie Chicken	897	111x	136	11,320	83x
Chimp	3,349	88x	72	9,969	138x
Human	3,086	39x	178	26,337	148x
Alligator	2,157	73x	81	21,540	265x



Importantly, Dovetail assemblies exhibit both high contiguity and high accuracy. Below is a table comparing misjoin rates and contiguity between two leading assemblers and Dovetail's HiRise pipeline for a human sample, NA12878. The ALLPATHS and Meraculous assemblies were generated with rich next generation sequencing datasets that included shotgun, mate-pair (MP), and fosmid data, while the HiRise assembly used only the same shotgun data and Chicago data. While ALLPATHS produces an impressive N50, it does so at a high cost in misjoins. Meraculous, designed to favor accuracy over contiguity, does much better with errors but does not perform as well in terms of contiguity. Dovetail's HiRise assembly achieves the best of both worlds, with industry-leading contiguity and accuracy.

Table 2: A comparison of Dovetail's scaffolding pipeline versus community assemblers.

Assembler	Input Data	Genome in scaffolds with misjoins	N50	Completeness
ALLPATHS	Shotgun, MP, fosmids	24.5%	12.1 Mb	92.2%
Meraculous	Shotgun, MP, fosmids	3.2%	9.1 Mb	94.8%
HiRise	Shotgun, Chicago	2.8%	12.6 Mb	94.1%

www.Dovetail-Genomics.com