De novo identification of "heterotigs" towards accurate and in-phase assembly of complex plant genomes

Jared C. Price[1], Joshua A. Udall[2], Paul M. Bodily[1], Judson A. Ward[3], Michael C. Schatz[4], Justin T. Page[2], James D. Jensen[1], Quinn O. Snell[1], and Mark J. Clement[1]

[1]Computer Science Department, Brigham Young University, Provo, UT, USA
[2]Plant and Wildlife Sciences Department, Brigham Young University, Provo, UT, USA
[3]Department of Horticulture, Cornell University, NYSAES, Geneva, New York, USA
[4]Simons Center for Quantitative Biology, CSHL, Cold Spring Harbor, New York, USA

Abstract—*Accurate and in-phase de novo assembly of highly polymorphic diploid and polyploid plant genomes remains a critical yet unsolved problem. "Out-of-the-box" assemblies on such data can produce numerous small contigs, at lower than expected coverage, which are hypothesized to represent sequences that are not uniformly present on all copies of a homologous set of chromosomes. Such "heterotigs" are not routinely identified in current assembly algorithms and could be used for haplotype phasing and other assembly improvements for such genomes. We introduce an algorithm which attempts to robustly identify heterotigs present in the assembly of a highly polymorphic diploid organism. The algorithm presented is for use with the 454 platform and for diploid assembly, but is readily adaptable to other sequencing platforms and to polyploid assembly.*

Keywords: heterozygous, genome, assembly, plant, heterotig, raspberry

1. Introduction

1.1 Background

Genome assembly is a relatively young field, but one which has been the subject of intense research. Motivated by a desire to reconstruct the human genome as rapidly as possible, the Whole Genome Shotgun strategy for genome assembly was introduced [1]. In this approach, genome structure inference is left entirely to software which takes as input a huge number of short DNA sequences ("reads") sampled from the entire genome. Although this approach was initially met with skepticism, a seminal paper provided the necessary proof of concept [2] and, due to its simplicity and cost-effectiveness, this approach has dominated genome projects since.

There are two primary classes of algorithms that are applied today to the Whole Genome Shotgun assembly problem. The first approach is referred to as the "overlap-layout-consensus" approach and the second approach is based on De Bruijn graphs. See [3] for a comparison of the two. We will focus on the overlap-layout-consensus approach, but the ideas regarding identification of heterotigs are applicable to both.

Overlap-layout-consensus assemblers often construct a data structure known as a "contig graph". A contig is simply a contiguous sequence of nucleotides inferred, via alignment of the input reads, to be present in the target genome. For various reasons, but primarily because of repetitive sequence, these contigs can essentially never be extended to full chromosome length in reasonably complex genomes, using current technologies. For this reason, the contig graph must represent not only the contigs themselves but also all of the possible adjacency relationships between contigs that are supported by the alignments. A common approach to representing this information, and the approach used in the 454 software, is to let the vertices of the graph represent contigs and the edges represent adjacency relationships between contigs. Because contigs have polarity (a 5' and a 3' end) the edges do not directly connect contigs, per se, rather, they connect specific ends of contigs. For example, an edge may indicate that the 5' end of contig 25 is adjacent to the 3' end of contig 1.

Critical to the upcoming discussion is a clear understanding of why assembly algorithms tend to collapse repetitive sequence into a single contig and the effect this has on the contig graph. Consider the case where a sequence of nucleotides (longer than the read length) occurs in the genome 3 times. Reads which are sampled from entirely within this repetitive sequence will align to each other with near perfect identity and will likely be collapsed into a single contig (in the absence of paired-end reads which align to unique sequences bordering the repeat). We will assume for demonstrative purposes that the sequences adjacent to each of the 3 copies are themselves unique. In the contig graph, the 5' end of the repeat contig will be adjacent to 3 different contigs, as will the 3' end (see Figure 1).

Notice that in Figure 1, in order to extend the contig that is currently represented as a collapsed 3-copy repeat any farther than the repeat sequence itself, you must accurately select a particular pair of contigs (one adjacent to the 5'

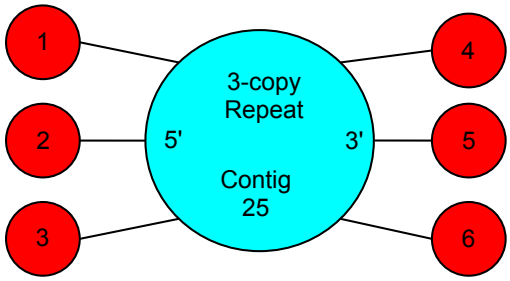

Fig. 1: A 3-copy repeat (collapsed in the assembly into a single contig) in a contig graph. Each circle is a vertex in the contig graph representing a particular contig in the assembly. Solid lines between contigs (more exactly between specific ends of those contigs) suggest that the contigs are adjacent to one another in the genome. Contigs 1-6 are single-copy contigs, each of which is adjacent to one of the 3 copies of the repeat. Each instance of the repeat is surrounded by a pair of contigs (one from the set $\{1, 2, 3\}$ and one from the set $\{4, 5, 6\}$). The pair of contigs surrounding a particular instance of the repeat constitute the "context" of that instance. When an assembly algorithm is unable to accurately determine the context around a particular copy of the repeat, contig extension must end at the repeat boundary. Worse, if the algorithm extends a contig through the repeat, but with the incorrect context, the resulting contig will contain sequence from 2 different locations in the genome.

end of the repeat and the other adjacent to the 3' end) with which to extend the contig. If you are not careful you might select contigs to use for the extension that are adjacent to different copies of the repeat in the actual genome, thereby constructing a contig that doesn't actually exist in the genome and whose 5' and 3' ends are in different locations in the genome. For this reason, repeats longer than the read length produce fragmentation of the contig graph and consequently smaller contigs in the assembly. The correct "context" for each copy of the repeat must be constructed carefully, usually using paired-end reads at a known distance and orientation with respect to each other.

1.2 Motivation

Highly polymorphic diploid and polyploid plant genomes have proven to be particularly difficult to assemble. Plants tolerate hybridization and polyploidization much more readily than most organisms that have been assembled by the Whole Genome Shotgun approach. These data present different challenges to assembly algorithms than those presented by highly homozygous diploid or monoploid organisms, for which traditional genome assembly algorithms are primarily designed. Notable examples of recent plant genome assembly projects include the small Fragaria vesca genome [4], a relatively heterozygous grapevine variety [5] and the large and ultra-repetitive maize genome [6].

Rubus idaeus cultivar 'Heritage' is an important commercial variety of raspberry which holds both biological and economic interest. Heritage is resistant to many of the most common raspberry diseases and has two raspberry subspecies in its recent pedigree, namely, *Rubus idaeus ssp. strigosus* and *Rubus idaeus ssp. vulgatus*. Such a scenario is not unique to Heritage, and is very common in raspberry breeding. Furthermore, hybridization, in general, is relatively common among plants.

Despite being very similar in appearance, amenable to hybridization, and prominent in the pedigrees of many commercial varieties of raspberry, these two subspecies have historically been geographically isolated with *strigosus* being a North American variety, and *vulgatus* a Eurasian variety. Furthermore, despite both varieties often being labeled as subspecies of *Rubus idaeus* taxonomists currently favor classifying these organisms as two different species, namely *Rubus strigosus* and *Rubus idaeus*.

Until recently, and to a great extent even today, the genomes of diploid and polyploid organisms have been assembled and presented in a monoploid form. Such an approach minimizes sequencing cost (greater depth is often required by algorithms that attempt to perform true diploid or polyploid assembly) and increases algorithmic simplicity for such tasks as genome assembly, mapping reads to a reference, and viewing a genome in a genome browser. Despite these advantages, such an approach also has distinct disadvantages. For example, diploid assemblies can provide a more accurate depiction of sequence diversity within a pair of homologous chromosomes than simple mapping back to a monoploid reference can provide. This information can then be used to improve numerous downstream analyses.

Genome assemblers that provide only a monoploid representation of a diploid or polyploid organism often contain algorithms that obscure sequence diversity or, worse, produce sequence not actually present in the target genome. For example, sequence diversity can be hidden when an algorithm deals with polymorphic regions by simply selecting one of the possible paths and ignoring all other possibilities. In the context of a highly heterozygous genome, the monoploid representation of the assembly can often "jump" between different members of a homologous set of chromosomes. Worse, an assembler may deal with polymorphic regions by producing a single contig that is a composite of the polymorphic paths in the contig graph, thereby producing sequence that isn't actually present on any chromosome.

With the advent of next-generation sequencing technologies, the field of genome assembly is aggressively pursuing more accurate and comprehensive representations. The Broad Institute's ALLPATHS-LG [7] is a notable example which represents the genome as the assembler actually sees it, that is to say, as a graph, thereby maintaining important

information about sequence diversity that may otherwise have been lost. Another fascinating approach, published very recently, applies colored de Bruijn graphs to the genome assembly problem in an attempt to assemble multiple eukaryotic genomes simultaneously [8] and to handle polymorphism in a more disciplined way.

We introduce an algorithm for identifying contigs present in an assembly which represent sequences that are not uniformly present on all members of a homologous set of chromosomes. We have chosen to call such contigs "heterotigs", and their counterparts, which are present on all members of the set, "homotigs". The algorithm presented here leverages coverage statistics, adjacency patterns between contigs in a contig graph, and paired end reads to identify heterotigs present in the assembly of a highly heterozygous diploid organism, and has been designed for use with the 454 sequencing platform, but the concepts are readily adaptable to polyploid assembly and to other sequencing platforms. Robust identification of heterotigs enables differential treatment of such sequences within an assembly algorithm and presents opportunities for producing more accurate and more complete assemblies of highly polymorphic species.

2. Heterotig Identification

We are now in a position to more formally define the problem with which this paper is primarily concerned. Let R represent a whole-genome set of sequencing reads from a highly polymorphic diploid species. Let C represent the set of contigs produced by an assembly of R, parameterized so as to separate "heterotigs" as cleanly as possible. Let E represent the set of edges in the contig graph. Let M represent the set of all meta-data available about the assembly, for example, alignment depths for each contig, contig lengths, etc. Let H represent the set of contigs whose sequence is found on only one copy of a homologous pair of chromosomes. Given C, E, and M is it possible to determine H to within an acceptable degree of accuracy? We will use a whole-genome sequencing data set from the highly heterozygous diploid organism *Rubus idaeus* 'Heritage' throughout this section as an example data set.

2.1 Inference Based on Coverage Statistics

The first question that arises in the context of identifying heterotigs is whether the depth of the read alignment from which a particular contig is constructed can be reliably used to infer the number of times the nucleotide sequence that contig represents is likely to appear in the target genome.

Consider the idealized case where read sampling from the genome is truly random and there are no other sources of coverage bias, for example from PCR artifacts or cloning bias. This idealized scenario is never realized in practice but is instructive for the real-world case which we will shortly turn to. Consider further that the organism being sequenced is diploid and expected to have very high rates of polymorphism. At every base in a particular contig there is a multiple alignment depth. Take the average of these depths across all bases in the contig and record this value as the "contig alignment depth".

What might the probability density function of contig alignment depths in a highly polymorphic diploid assembly look like? Let's say for illustrative purposes that we have sequenced the genome to 60x coverage, which is now routinely done with the advent of next generation sequencing. For a diploid organism, genome coverage is typically calculated in terms of the haploid genome size (total number of bases / haploid genome size), so this number is equivalent to the coverage we should expect for a single-copy homotig. We expect single-copy homotigs to be numerous and therefore expect a mode at approximately 60 in the probability density function. By this same logic, if heterotigs are indeed present in the assembly in significant amounts a mode should also be present at about half that coverage (30x). We expect there to be some breadth to the distribution around each peak and so high coverage will likely be necessary to determine if the modes are indeed present. Some of the density will be at much higher coverage (high-copy-number repeats) but we probably have no reason to expect that a particular copy number is more prevalent than another for high-copy-number repeats, so we expect no significant modes above our single-copy homotig mode.

Let's now turn our attention to a real-world case. A recent whole-genome shotgun assembly project collected high-coverage sequence data from *Rubus idaeus* cultivar 'Heritage'. The sequence was assembled using the 454 assembler and the resulting contigs were queried for their contig alignment depths (see Figure 2).

Close examination of Figure 2 illuminates several interesting properties of the contigs from this assembly. First, and most obviously, modes corresponding to our theoretical peaks (one peak composed primarily of single-copy heterotigs and another peak composed primarily of single-copy homotigs) are clearly discernible across contigs of all lengths. If these peaks represent what we have hypothesized, the homotig-mode to heterotig-mode ratio should be very near 2, as is indeed the case, with the value ranging between 2.05 and 2.17 for the sets of contigs examined. Could there be another explanation besides the heterotig-homotig hypothesis we have presented for the strongly bimodal distribution? If so, the alternate hypothesis must account for why the lower mode (lower in terms of the coverage value, not necessarily peak height) is at nearly exactly half the coverage of the higher mode.

More encouraging (for the purposes of heterotig identification) than the mere presence of the peaks is the observation that for many of the sets of contigs examined the density between the peaks is very low, suggesting that, at least for this data set, coverage can be used to make inference on copy

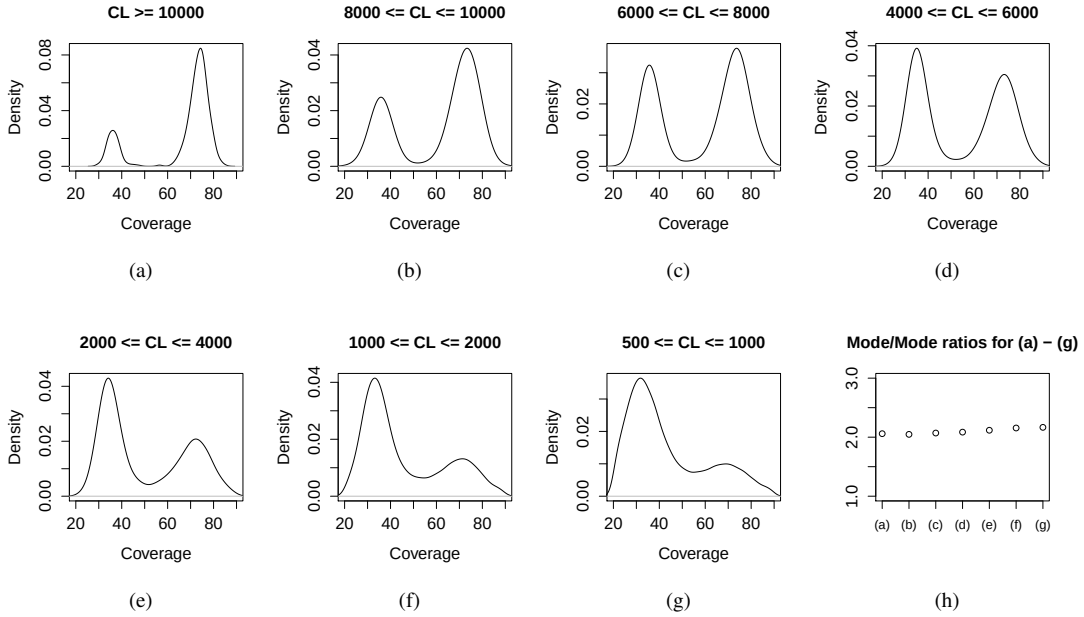

Fig. 2: (a)-(g) Probability density functions (PDFs) of contig alignment depth calculated from the set of contigs produced in an assembly of *Rubus idaeus* 'Heritage'. Contig alignment depth is defined as the mean of the single-position alignment depths calculated at each position in the contig. Each PDF analyzes contigs within a particular length class (CL = Contig Length). Contigs with contig alignment depths outside of the interval $[20, 90]$ are excluded. The largest contigs are predominantly at "homotig" coverage while the smaller contigs are predominantly at "heterotig" coverage. (h) "Homotig" peak mode over "heterotig" peak mode ratios for (a)-(g). The minimum value was 2.05 and the maximum value was 2.17

number. The bimodal nature of the distribution is consistent across contigs of all sizes. In contrast, the relative density under each peak differs dramatically for contigs in different length classes. The longest contigs are predominantly single-copy homotigs while the shorter contigs are predominantly single-copy heterotigs. Furthermore, as the contig length gets smaller the density between the peaks increases, although never enough to make the peaks difficult to see.

2.2 Inference Based on Contig Graph Structure

If our hypothesis from the previous section is accurate, namely, that the bimodal PDFs in the previous section suggest an extremely heterozygous diploid genome where many of the contigs are present on only a single chromosome (as opposed to both chromosomes of a homologous pair), then it is safe to assume that many of the heterotigs will be broken at boundaries where they are adjacent to single-copy homotigs. Consider a chromosome A and its homologous pair B. Now consider a single-copy homotig C that is present on both A and B. On chromosome A, C is adjacent to a single-copy heterotig D. On chromosome B, by the definition of heterotig, C must be adjacent to some sequence other than D, and consequently, the extension of contig C must be broken to account for these 2 different adjacencies. Recalling that assemblers must break contigs whenever there is a repeat longer than the read length (see Figure 1), notice that in the context of such extreme heterozygosity, single-copy homotigs can behave similarly to 2-copy repeats, having one context in one homologous chromosome and another context in the other, providing one explanation for the extremely bimodal PDFs presented in the previous section.

Assuming this explanation is correct, such data are not likely to assemble well using traditional assembly algorithms. First, the assembly is likely to be extremely fragmented, with thousands, if not hundreds of thousands, of small contigs. There will be many more "ambiguous" adjacency relationships between contigs than would be seen in either homozygous diploid or monoploid assemblies. Furthermore, the extent to which homotig order is consistent in the two members of a homologous pair is critical to the tractability of an algorithmic solution. If the order of single-copy homotigs is strictly consistent the problem is greatly simplified. Under that scenario, only a few different signature patterns should occur in the contig graph, for example, it is probably safe to assume that under such a condition the graph should contain numerous "bubbles", locations where a single-copy homotig bifurcates to two single-copy heterotigs which both immediately converge to a second single-copy homotig.

Length Class	Percentage
length >= 10000	78 %
8000 <= length <= 10000	76 %
6000 <= length <= 8000	78 %
(4000 <= length <= 6000	82 %
2000 <= length <= 4000	86 %
1000 <= length <= 2000	89 %
500 <= length <= 1000	90 %

(a)

Length Class	Percentage
length >= 10000	2 %
8000 <= length <= 10000	2 %
6000 <= length <= 8000	6 %
4000 <= length <= 6000	8 %
2000 <= length <= 4000	18 %
1000 <= length <= 2000	31 %
500 <= length <= 1000	41 %

(b)

Fig. 3: (a) All contigs of alignment depth between 25 and 40 in various length classes were marked as "candidate heterotigs". The percentages given indicate the percentage of candidate heterotigs connected on either the 5' or 3' end to at least one contig end which participated in exactly 2 edges (suggestive of a homotig-heterotig boundary possibly being the cause for contig breakage). (b) The same as (a) except the percentage now reflects the percentage of candidate heterotigs that were found in "perfect bubbles". See Figure 4 for the precise way in which we have defined the term "perfect bubble".

If this scenario predominates, assembling two homologous chromosomes exhibiting extremely high heterozygosity would, to a considerable extent, reduce to the problem of identifying heterotigs, and subsequently treating heterotig-to-heterotig paired-end data differently than homotig-to-homotig paired-end data. Homotig-to-homotig paired-end data would help lay out the structure shared between the two members of the pair and heterotig-to-heterotig paired-end data could help keep one chromosome separate from the other, to as great a degree as possible, when building contigs. Notice that the higher the rate of heterozygosity in this scenario the better because it gives you more heterotig anchors for keeping each chromosome "in phase".

What about the case where the order and orientation of the homotigs differs somewhat between homologs? This would mean that in addition to assembly "bimodality" in the sense of having significant populations of both heterotigs and homotigs, there would also be assembly bimodality in the relationships between homotigs (a certain set of relationships prevailing on one homolog, and another set of relationships prevailing on the other). For example, on one chromosome, a pair of homotigs may occur at one distance and orientation with respect to each other, yet on the homolog, the same pair of homotigs may occur at a different distance and/or orientation. Such a scenario would obviously pose tremendous difficulties to traditional genome assembly algorithms. How do you correctly estimate the singular distance between two homotigs using paired end data when there are, in fact, two distances? How do you layout a genome when there are, in fact, two different layouts? The problems posed in this scenario would require the assembler itself to also be "bimodal", that is to say, it would have to deal differentially with each homolog. The multiple "modes" could be represented using multiple graphs or by having multiple passes through the same graph. In either case, the assembler would need robust and accurate identification of heterotigs throughout the process.

The current manuscript does not attempt to perform a comprehensive analysis of the contig graph patterns observed in the assembly of *Rubus idaeus* 'Heritage', however, Figure 3 provides a sense of what the contig graph looks like internally. In particular it examines what the contig graph looks like immediately around "candidate heterotigs" (contigs that appear to be heterotigs based on coverage alone). Notice that for contigs in every length class examined, large majorities of the candidate heterotigs are connected either on their 5' or 3' end to a contig end that participates in exactly two edges, providing a measure of supporting evidence for a homotig-heterotig boundary (a particular end of a single-copy homotig, which is adjacent to a heterotig in one homolog, would likely be adjacent to exactly one other sequence in the other homolog, thereby participating in exactly 2 edges). Furthermore, only a relatively small percentage of heterotigs are found in "perfect bubble" patterns in the contig graph, suggesting that algorithms which rely on simple graph patterns to identify heterotigs may significantly underestimate the true sequence diversity. It is also interesting that, as the average length of a set of candidate heterotigs decreases, the percentage of those candidate heterotigs found in perfect bubbles increases (see Figure 3).

3. Algorithm

Definitions:

$A =$ Alignment depth of a contig
$B_{hc} =$ Boolean, true if $H_{min} \leq A \leq H_{max}$
$B_{per} =$ Boolean, true if H_{cand} is in a perfect bubble
$C =$ A contig (a node from G_{454})
$C_e =$ A contig end (5' or 3')
$C_{num} =$ The total number of contigs in the assembly
$G_{454} =$ a 454ContigGraph.txt file (from Newbler)
$H =$ The true set of heterotigs
$H_c = \{C : C \in H\}$ (with high confidence)
$H_{cand} =$ Any C where B_{hc} holds

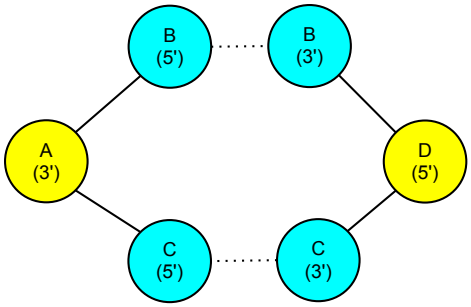

Fig. 4: Graphical depiction of a perfect bubble in a contig graph. Edges connecting contig ends are denoted with solid lines. A, B, C, and D identify contigs. (5') and (3') each identify a particular end of a contig. We say the structure is a perfect bubble when the following hold: (1) A, B, C, and D are 4 distinct contigs. (2) All 4 ends of the heterotigs (B and C) participate in exactly one edge each. (3) The ends of A and D that are connected to the heterotigs participate in exactly 2 edges each.

H_{max} = Maximum A for a heterotig candidate
H_{min} = Minimum A for a heterotig candidate
L = Length of a contig
P = The set of all "paired-end flows" reported in G_{454}

Domain: $\{x : x = G_{454}\}$
Range: $\{y : y = H_c\}$

function IDENTIFYHETEROTIGS(H_{min}, H_{max})
  Add to H_c *all* H_{cand} *such that* B_{per} *holds*
  for *all* H_{cand} *with* $L \geq 2000$ **do**
    if H_{cand} *connects to bifurcating* C_e **then**
      Add H_{cand} *to* H_c
    end if
  end for
  while H_c *grows with each iteration* **do**
    for *all* H_{cand} **do**
      if P *links* H_{cand} *with* $e \in H_c$ **then**
        Add H_{cand} *to* H_c
      end if
    end for
  end while
return H_c
end function

4. Discussion

We have presented evidence that complex plant genomes, particularly highly heterozygous organisms arising through hybridization or polyploidy, present unique and difficult challenges to the Whole Genome Shotgun assembly problem that are not encountered in either monoploid or homozygous genome assembly.

When heterozygosity rates are sufficiently high, and coverage sufficiently deep, it is possible to perform de novo identification of "heterotigs" (sequences not uniformly present on all copies of a homologous set of chromosomes) via inference on a combination of coverage statistics, contig graph patterns, and paired end reads (when available). These heterotigs can then serve as guideposts in the assembly process to improve assembly quality and completeness, as well as to minimize how often the assembled scaffolds and contigs "jump" from sequence in one homolog to sequence in the other.

We have also given preliminary evidence suggesting that algorithms that identify heterotigs via very simple graph patterns, such as the perfect bubbles analyzed in section 2.2, are likely to underestimate true sequence diversity in highly heterozygous species. Furthermore, we have suggested several ways in which more robust identification of heterotigs could lead to more accurate and complete assemblies for such data. This scenario necessitates a more rigorous treatment of "heterotigs" which we begin laying the foundation for here.

We believe that robust identification of, and intelligent treatment of, such sequences could dramatically improve the state of the art with regards to the genome assembly of highly polymorphic diploid and polyploid species.

References

[1] J. L. Weber and E. W. Myers, "Human whole-genome shotgun sequencing," *Genome Research*, vol. 7, no. 5, pp. 401–409, 1997. [Online]. Available: http://genome.cshlp.org/content/7/5/401.short

[2] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000. [Online]. Available: http://www.sciencemag.org/content/287/5461/2196.abstract

[3] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan, "Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph," *Briefings in Functional Genomics*, 2011. [Online]. Available: http://bfg.oxfordjournals.org/content/early/2011/12/18/bfgp.elr035.abstract

[4] V. Shulaev, D. Sargent, R. Crowhurst, T. Mockler, O. Folkerts, A. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. Mane, *et al.*, "The genome of woodland strawberry (fragaria vesca)," *Nature genetics*, vol. 43, no. 2, pp. 109–116, 2010.

[5] R. Velasco, A. Zharkikh, M. Troggio, D. Cartwright, A. Cestaro, D. Pruss, M. Pindo, L. FitzGerald, S. Vezzulli, J. Reid, *et al.*, "A high quality draft consensus sequence of the genome of a heterozygous grapevine variety," *PLoS One*, vol. 2, no. 12, p. e1326, 2007.

[6] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges,

E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, "The b73 maize genome: Complexity, diversity, and dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009. [Online]. Available: http://www.sciencemag.org/content/326/5956/1112.abstract

[7] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe, "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences*, vol. 108, no. 4, pp. 1513–1518, 2011. [Online]. Available: http://www.pnas.org/content/108/4/1513.abstract

[8] Z. Iqbal, M. Caccamo, I. Turner, P. Flicek, and G. McVean, "De novo assembly and genotyping of variants using colored de bruijn graphs," *Nature Genetics*, 2012.