

Analysis of copy number variation in *C. elegans* using next generation sequencing

B. Kingham¹, B. Sanderson², A. Noll², H. Escobar³, S. Blake⁴, K. Jonscher⁵, A. Hutchinson⁶, C. Dagnall⁶, C. Lytle⁷

1) University of Delaware, Newark, DE. 2) Stowers Institute for Medical Research, Kansas City, MO. 3) Eurofins MWG Operon, Huntsville, AL. 4) University of Missouri, Columbia, MO. 5) University of Colorado, Denver, CO. 6) CGF, NCI, SAIC-Frederick, Gaithersburg, MD. 7) Dartmouth College, Hanover, NH.

Abstract

Structural variations (SV) found in eukaryotic genomes include insertions, deletions, inversions, translocations, and copy number variations (CNV). The emerging body of literature clearly illustrates the role of SVs, such as CNVs, in the susceptibility or resistance to certain diseases. While microarrays have traditionally been an effective tool for identifying CNVs, recent advances in mate-pair, next-generation sequencing technology now provide an alternative approach.

To detect copy number variants using mate-pair sequencing it is essential to leverage a bioinformatics pipeline that can distinguish the mapped ends, and identify statistically significant differences as compared to a reference sequence. Several commercial and open source software tools for automatic detection of SVs and CNVs are available and this list is rapidly growing. Though the number of tools continues to increase, they are neither as robust nor mature as those currently available for analyzing microarray experiments. As such, packages are being constantly evaluated with the intent of determining which performs best in this capacity.

For its annual research project, the GVRG has hypothesized that an optimal combination of the statistical models and paired-end reads will have the most traction in next-generation sequencing for CNV detection. Using *C. elegans* as a model organism, we have performed and directed experiments to study the capability of next-generation sequencing for such a purpose. It has been reported that there is nearly 2% natural gene content variation between the Bristol and Hawaii *C. elegans* strains as determined by aCGH. These published differences, which include a number of CNVs, have provided a valuable framework for conducting the experiments and analyzing results.

Copy Number Variants

Definition:
In an effort to standardize the experimental setup and analysis of our study, we sought to clearly define CNVs. Using a 2007 paper published by Nature Genetics as a guideline, we describe a CNV as "a DNA segment of at least 1 kb in size, for which copy number differences have been observed in the comparison of two or more genomes." Under its simplest definition, the term "carries no implication of relative frequency or phenotypic effect".

Biological effect of CNVs:

- CNVs are the largest class of known structural variants
- Estimated that 0.4% of the genomes of unrelated humans differ with respect to copy number
- CNVs can be used to predict metastatic capabilities of cancers, have been shown to be prevalent in most cancers and are implicated in adaptive evolution processes such as emergence of drug resistance.
- They are associated with numerous diseases including autism, schizophrenia, Alzheimers, Parkinsons, early onset obesity and susceptibility to HIV.

array Comparative Genome Hybridization (aCGH)

As a foundation for this study, we used an aCGH publication where the authors employed a microarray to interrogate CNVs between two strains of *C. elegans*. This array consisted of ~400K, 50-mer probes tiled across the exons of all 6 *C. elegans* chromosomes.

Methods

Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization

Jason S. Maydan,¹ Stephane Flibotte,² Mark L. Edgley,³ Joanne Lau,³ Rebecca R. Selzer,⁵ Todd A. Richmond,⁵ Nathan J. Pofahl,⁵ James H. Thomas,⁴ and Donald G. Moerman,^{1,3,6}

¹Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; ²Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 4S6 Canada; ³Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; ⁴Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; ⁵NimbleGen Systems Inc., Madison, Wisconsin 53711, USA

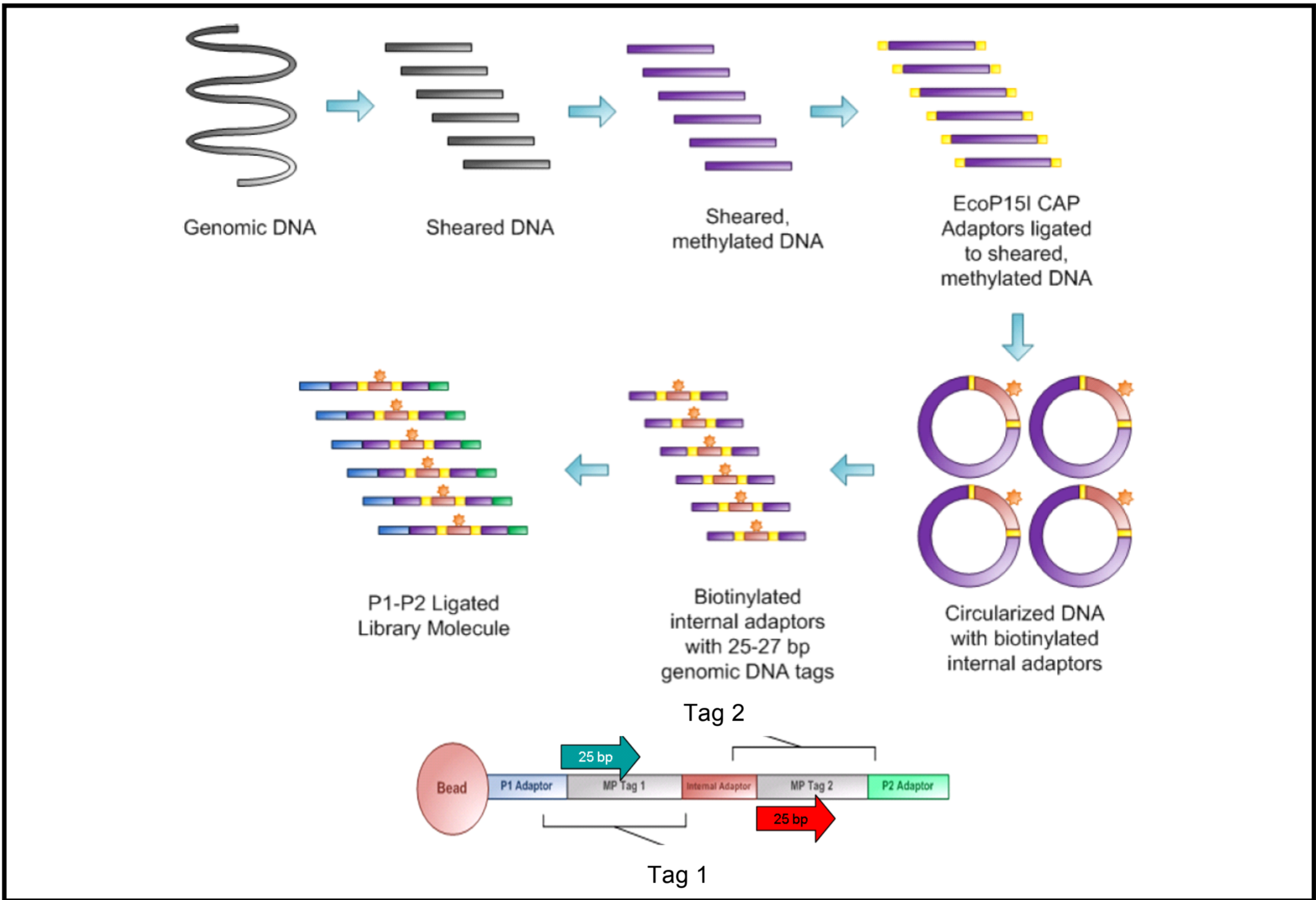
The authors report a nearly 2% difference in Hawaii vs Bristol strains of *C. elegans* due to copy number variants. In addition, they provide the name and genomic location for 531 of these CNVs. These variants are the basis for our study.

Methods

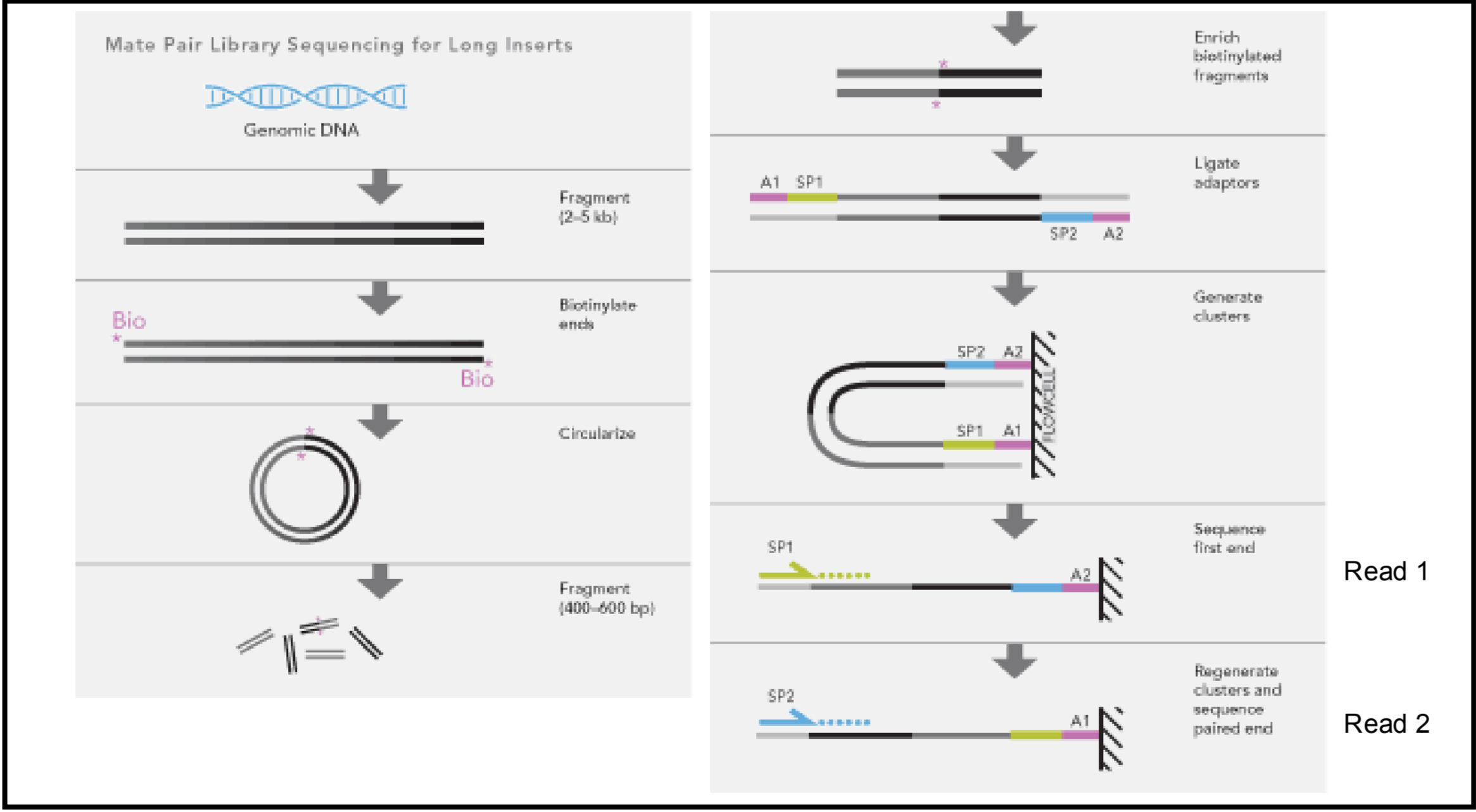
***C. elegans* gDNA preparation:**
An isogenic population of the *C. elegans* Hawaii strain was grown on *E. coli*. Genomic DNA was prepared from this population using a phenol/chloroform-based extraction technique adapted from Current Protocols in Molecular Biology (2003, Unit 13.11). The *C. elegans* genome consists of 5 pairs of autosomes along with 1 pair of sex chromosomes and contains ~ 100 Mbp of DNA.

Next gen-sequencing:
Genomic DNA was sent to Life Technologies for mate-paired library construction and sequencing on the ABI SOLiD 3 Plus. Genomic DNA was sent to Illumina for mate-paired library construction and sequencing on the Illumina Genome Analyzer IIx. Mate pair library size was left to the discretion of each company. Life Technologies constructed a 1.5Kb mate pair library and Illumina constructed 2Kb, 2.5Kb, 3Kb and 4Kb libraries for sequencing. The SOLiD instrument provided 42.22 G bases of sequence corresponding to ~180X mean depth and the GAIi gave 14.17 G bases for a total of ~76X mean depth.

SOLiD mate-pair library construction



GAIi mate-pair library construction



Alignment of Reads

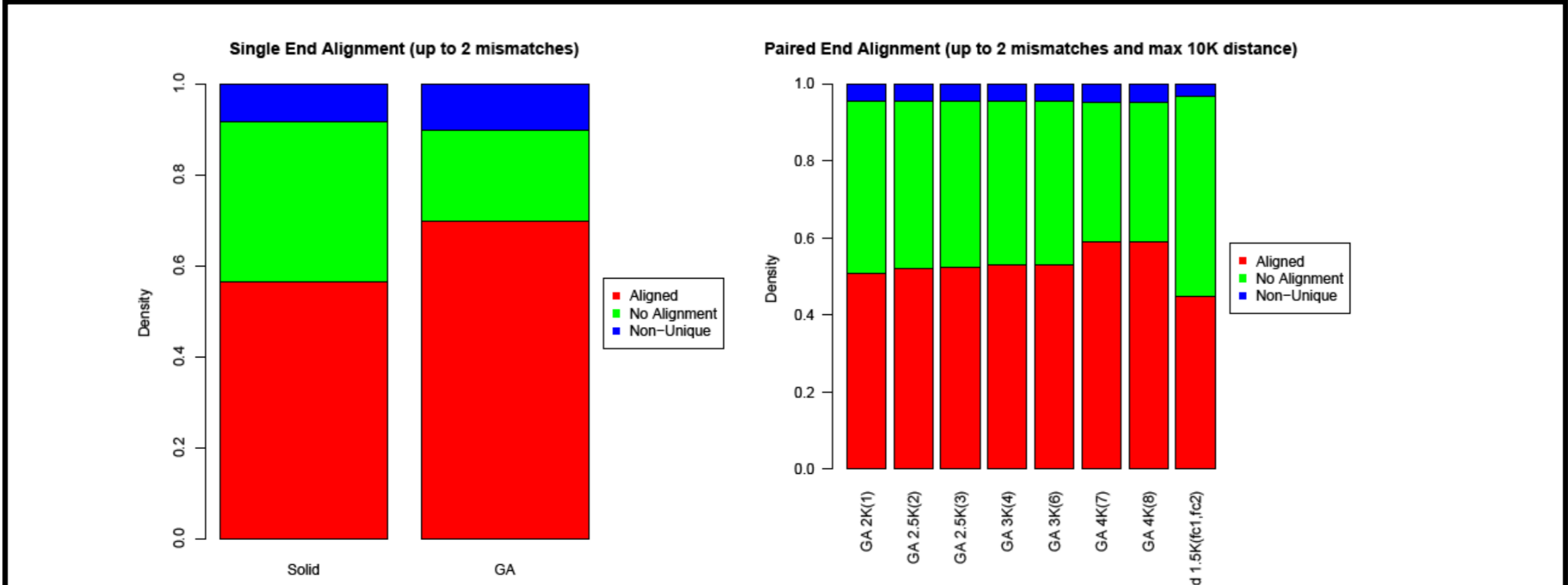


Figure 1. Reads were aligned using bowtie. Alignments were required to be unique, allowing up to two mismatches per read and a 10Kb distance between pairs.

Methods

End Sequence Profiling (ESP):
Many different types of mapping signatures may be encountered when analyzing ESP data. The most basic scenarios of simple insertions, deletions and concordance are depicted in figures 1-3 below.

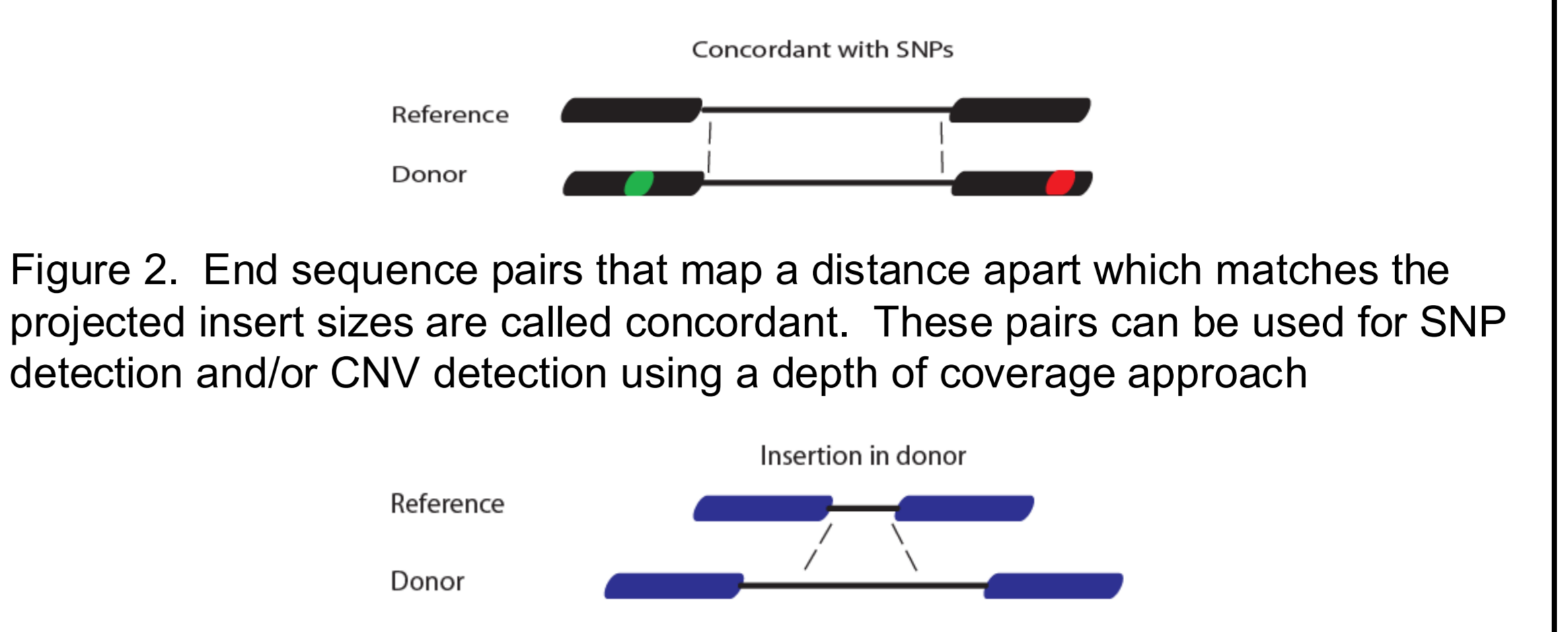


Figure 2. End sequence pairs that map a distance apart which matches the projected insert sizes are called concordant. These pairs can be used for SNP detection and/or CNV detection using a depth of coverage approach

Figure 3. An insertion can be detected when the distance between the locations where end sequences map in the reference genome is smaller than the expected library insert size. In ESP, the maximum detectable donor insertions size is limited by the library size insert.

Figure 4. A deletion can be detected when the distance between the locations where end sequences map in the reference genome is greater than the expected library insert size.

Depth of Coverage (DOC):
HTS allows for the detection of gain/loss events by examining the levels of aligned reads throughout the genome. Both paired and un-paired reads can be used for detecting CNVs with the DOC measure.



Figure 5. A potential duplication signature where the number of reads mapped to a given location is much greater than expected given the total number of reads and genome size



Figure 6. A potential deletion signature where the number of reads mapped to a given location is zero or much lower than expected given the total number of reads and genome size

Split Mapping:
Indels can be contained internally within reads from a donor genome where the beginning and end of a read map to different areas of the genome. Due to the short reads of HTS, many false positive split mappings may occur. By assuming one end of a pair must map a given distance from the other, the false positive split mappings can be reduced. In addition, clusters of anchored split mappings can be used to increase the confidence of a prediction.

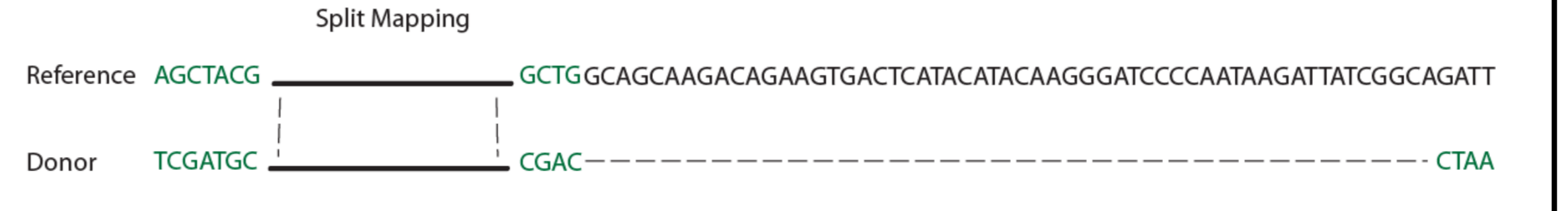
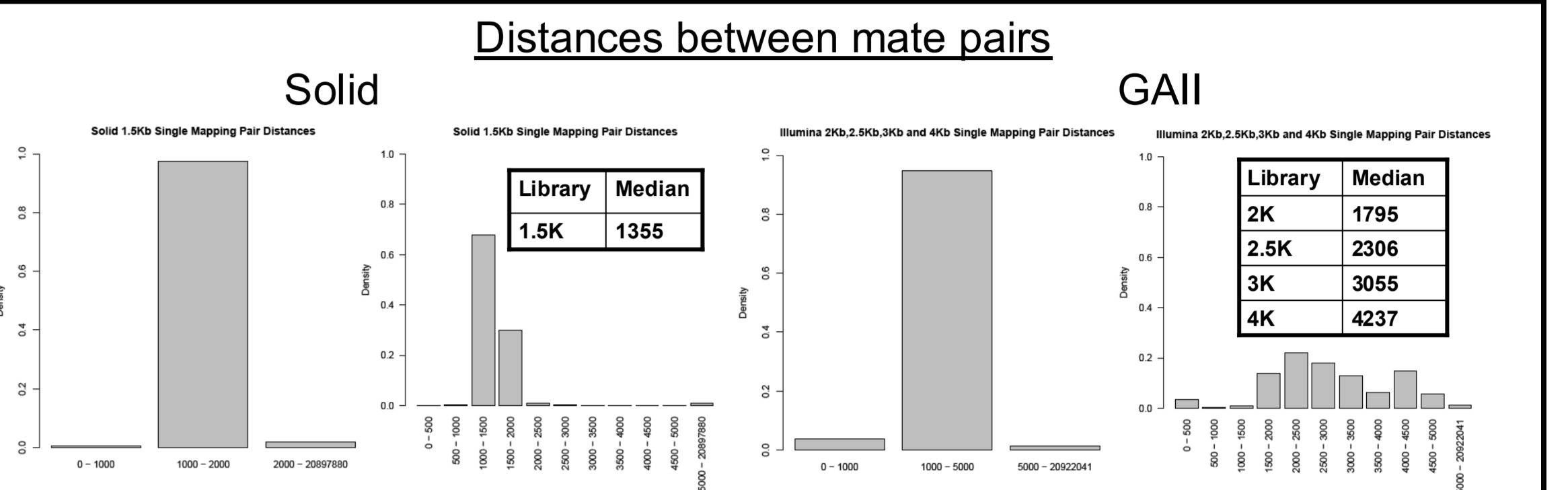


Figure 7. A deletion can be located within one end of a read pair by leveraging an alignment for the other end.

Results



CNVs found by Solid and GAIi using ESP compared to aCGH



Figure 8. Using Breakdancer on the Solid and GAIi data, 409 of 531 Maydan CNVs were found. Figure 9. The Solid CNV DOC tool found an additional 100 CNVs

Future Directions - CNV Validation

Upon completion of our primary analysis, we will pick a number of CNVs to validate. These CNVs will consist of a subset of true positives (known CNV region detected by sequencing), false negatives (known CNV region not detected by sequencing), and false positives or novel CNVs (new CNV regions detected by sequencing).

We will utilize LifeTechnology's CopyCaller workflow to validate the CNVs detected. This technique uses qPCR along with TaqMan[®] Copy Number assays to quantify the number of copies of a specific region. Both the Hawaiian and Bristol strains will be assessed for the CNVs of interest during the validation procedure.

Future Directions - Structural Variants

We will make use of our dataset to look at structural variants including inversions, translocations, and other genomic permutations.

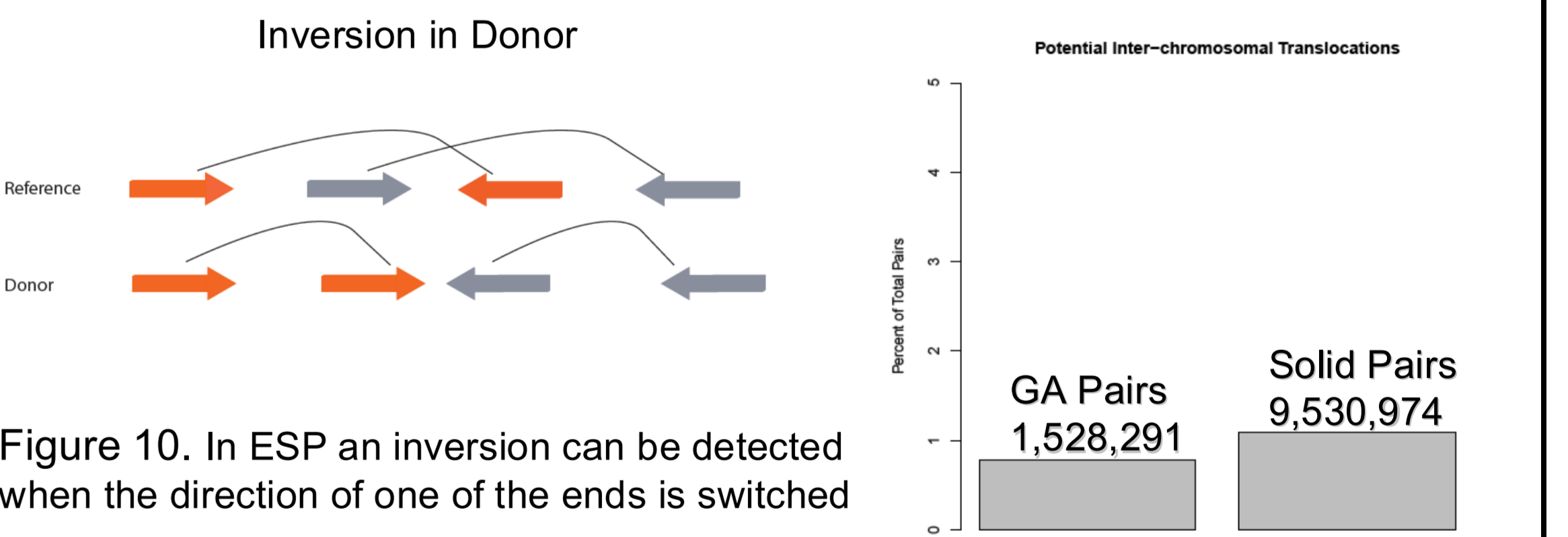
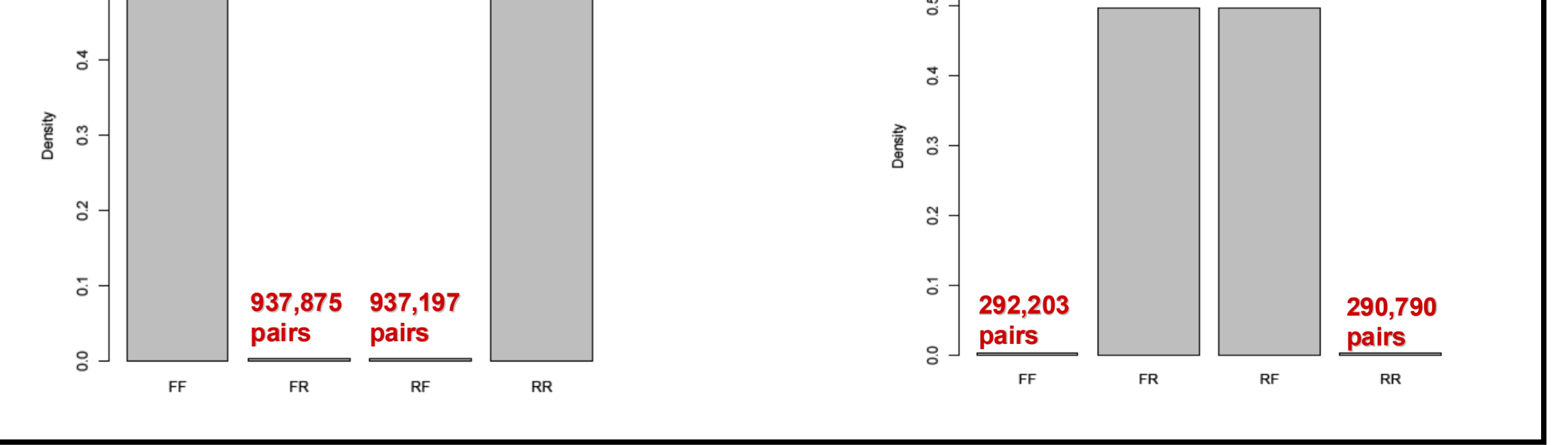


Figure 10. In ESP an inversion can be detected when the direction of one of the ends is switched



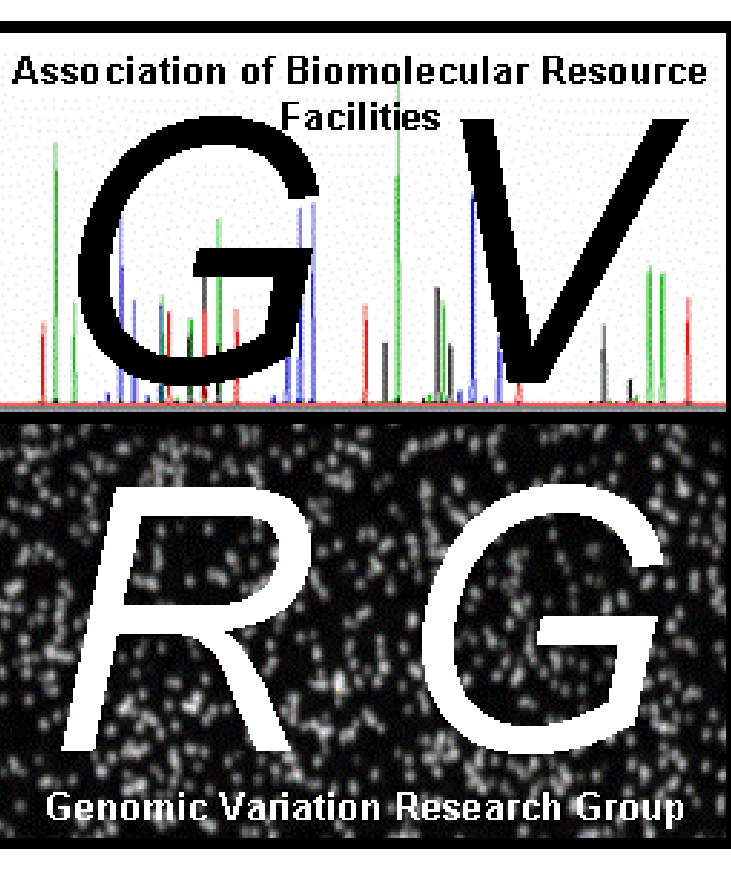
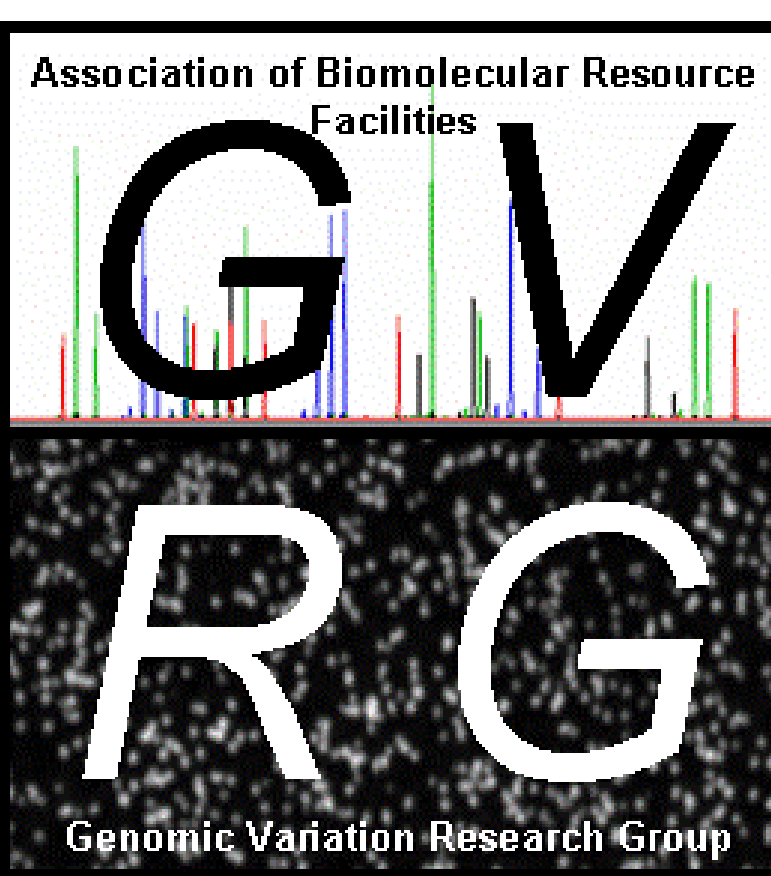
We would like to thank

- Charles Cochran and Life Technologies for generating SOLiD data, expert data analysis and CNV validation using their CopyCaller workflow.
- Joel Fellis and Illumina for generation of Illumina GAIi data.
- The HoYi Mak lab at the Stowers Institute for providing the *C. elegans* strains
- Ken Chen et al. for Breakdancer

Analysis of copy number variation in *C. elegans* using next generation sequencing

B. Kingham¹, B. Sanderson², A. Noll², H. Escobar³, S. Blake⁴, K. Jonscher⁵, A. Hutchinson⁶, C. Dagnall⁶, C. Lytle⁷

1) University of Delaware, Newark, DE. 2) Stowers Institute for Medical Research, Kansas City, MO. 3) Eurofins MWG Operon, Huntsville, AL. 4) University of Missouri, Columbia, MO. 5) University of Colorado, Denver, CO. 6) CGF, NCI, SAIC-Frederick, Gaithersburg, MD. 7) Dartmouth College, Hanover, NH.



Abstract

Structural variations (SV) found in eukaryotic genomes include insertions, deletions, inversions, translocations, and copy number variations (CNV). The emerging body of literature clearly illustrates the role of SVs, such as CNVs, in the susceptibility or resistance to certain diseases. While microarrays have traditionally been an effective tool for identifying CNVs, recent advances in mate-pair, next-generation sequencing technology now provide an alternative approach. To detect copy number variants using mate-pair sequencing it is essential to leverage a bioinformatics pipeline that can distinguish the mapped ends, and identify statistically significant differences as compared to a reference sequence. Several commercial and open source software tools for automatic detection of SVs and CNVs are available and this list is rapidly growing. Though the number of tools continues to increase, they are neither as robust nor mature as those currently available for analyzing microarray experiments. As such, packages are being constantly evaluated with the intent of determining which performs best in this capacity. For its annual research project, the GVRG has hypothesized that an optimal combination of the statistical models and paired-end reads will have the most traction in next-generation sequencing for CNV detection. Using *C. elegans* as a model organism, we have performed and directed experiments to study the capability of next-generation sequencing for such a purpose. It has been reported that there is nearly 2% natural gene content variation between the Bristol and Hawaii *C. elegans* strains as determined by aCGH. These published differences, which include a number of CNVs, have provided a valuable framework for conducting the experiments and analyzing results.

Copy Number Variants

Definition:
In an effort to standardize the experimental setup and analysis of our study, we sought to clearly define CNVs. Using a 2007 paper published by Nature Genetics as a guideline, we describe a CNV as "a DNA segment of at least 1 kb in size, for which copy number differences have been observed in the comparison of two or more genomes." Under its simplest definition, the term "carries no implication of relative frequency or phenotypic effect".

Biological effect of CNVs:

- CNVs are the largest class of known structural variants
- Estimated that 0.4% of the genomes of unrelated humans differ with respect to copy number
- CNVs can be used to predict metastatic capabilities of cancers, have been shown to be prevalent in most cancers and are implicated in adaptive evolution processes such as emergence of drug resistance.
- They are associated with numerous diseases including autism, schizophrenia, Alzheimers, Parkinsons, early onset obesity and susceptibility to HIV.

array Comparative Genome Hybridization (aCGH)

As a foundation for this study, we used an aCGH publication where the authors employed a microarray to interrogate CNVs between two strains of *C. elegans*. This array consisted of ~400K, 50-mer probes tiled across the exons of all 6 *C. elegans* chromosomes.

Methods

Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization

Jason S. Maydan,¹ Stephane Flibotte,² Mark L. Edgley,³ Joanne Lau,³ Rebecca R. Selzer,⁵ Todd A. Richmond,⁵ Nathan J. Pofahl,⁵ James H. Thomas,⁴ and Donald G. Moerman,^{1,3,6}

¹Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; ²Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 4S6 Canada; ³Michael Smith Laboratories, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; ⁴Department of Genome Sciences, University of Washington, Seattle, Washington 98195-7730, USA; ⁵NimbleGen Systems Inc., Madison, Wisconsin 53711, USA

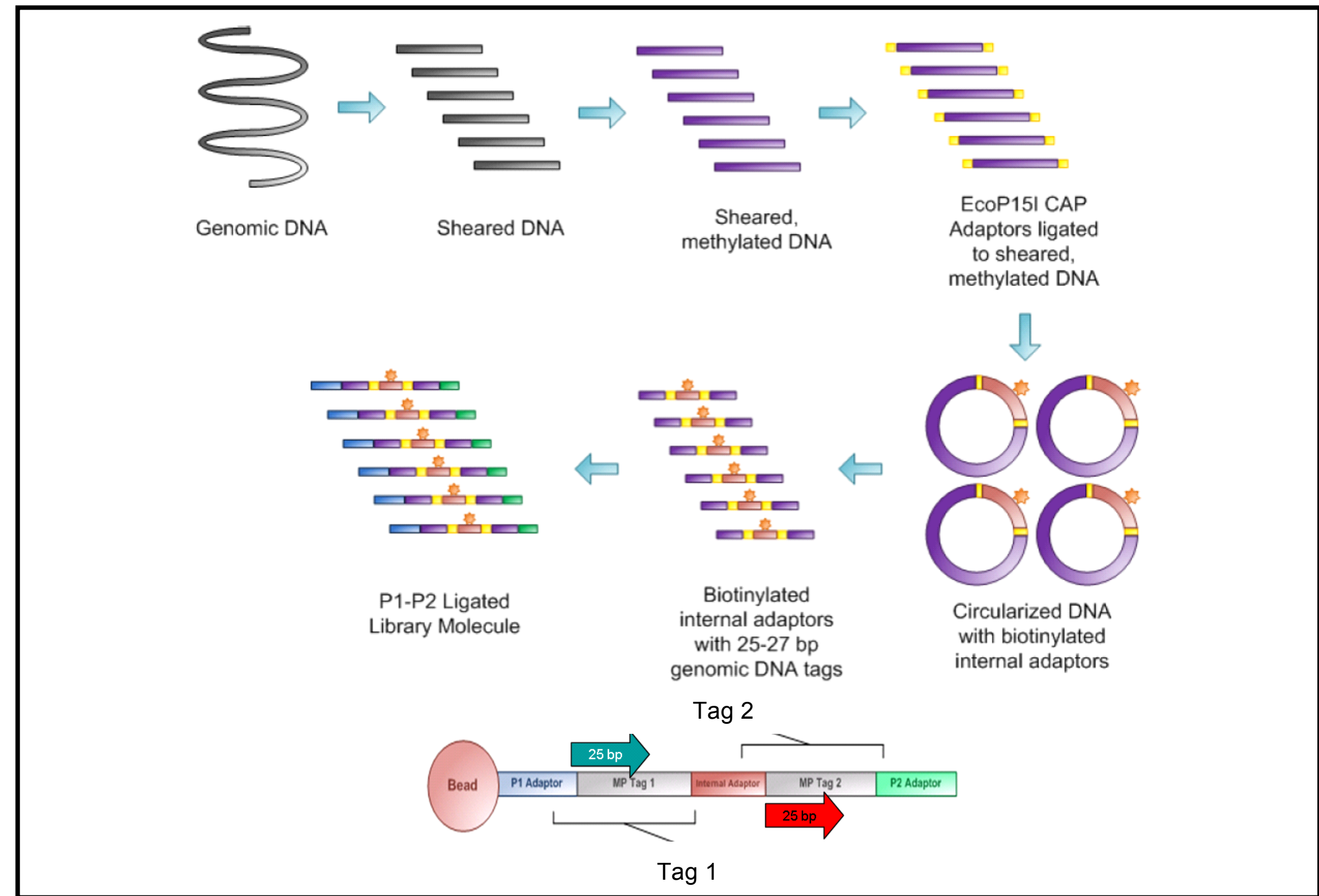
The authors report a nearly 2% difference in Hawaii vs Bristol strains of *C. elegans* due to copy number variants. In addition, they provide the name and genomic location for 531 of these CNVs. These variants are the basis for our study.

Methods

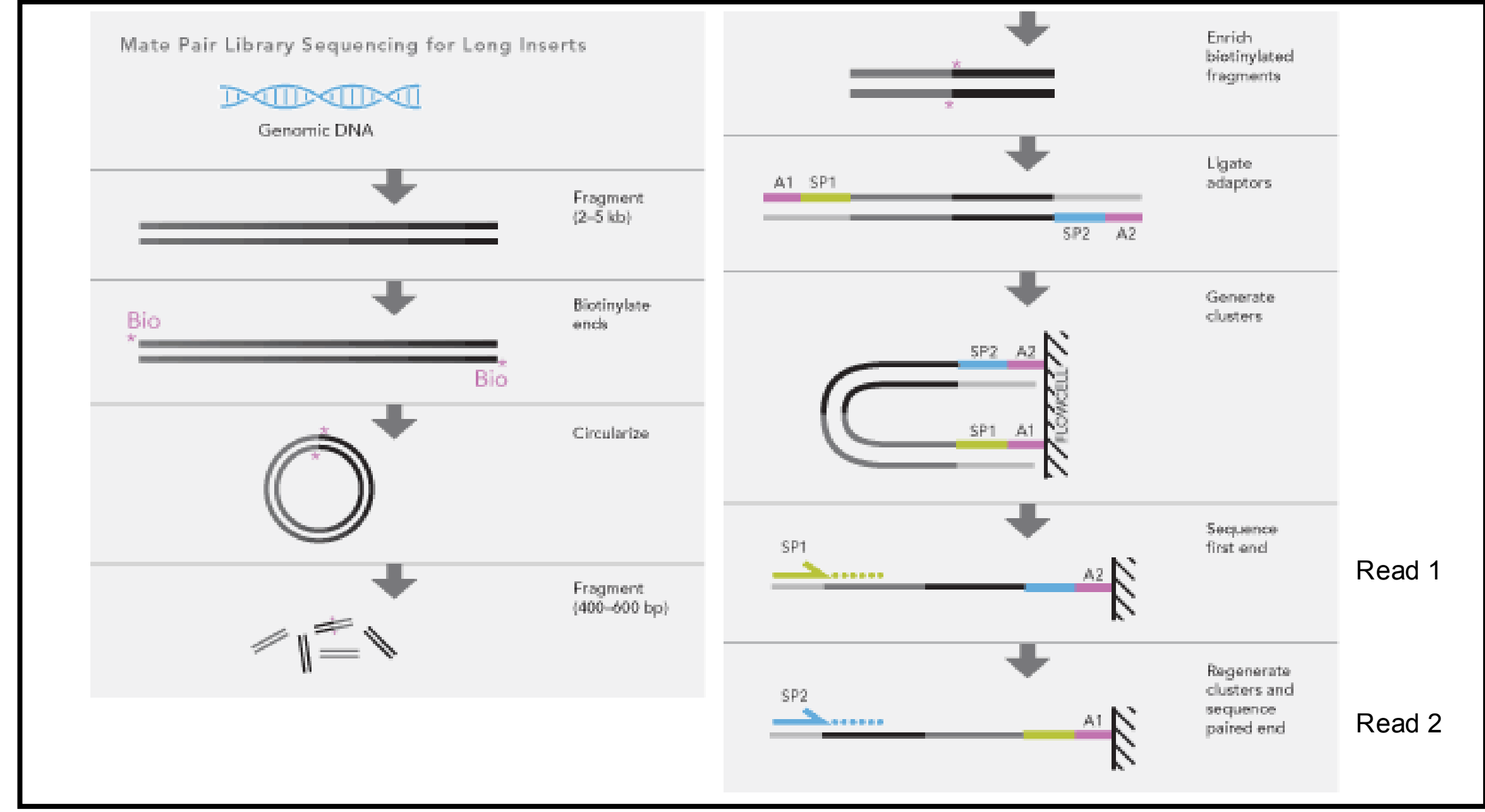
***C. elegans* gDNA preparation:**
An isogenic population of the *C. elegans* Hawaii strain was grown on *E. coli*. Genomic DNA was prepared from this population using a phenol/chloroform-based extraction technique adapted from Current Protocols in Molecular Biology (2003, Unit 13.11). The *C. elegans* genome consists of 5 pairs of autosomes along with 1 pair of sex chromosomes and contains ~ 100 Mbp of DNA.

Next gen-sequencing:
Genomic DNA was sent to Life Technologies for mate-paired library construction and sequencing on the ABI SOLiD 3 Plus. Genomic DNA was sent to Illumina for mate-paired library construction and sequencing on the Illumina Genome Analyzer IIx. Mate pair library size was left to the discretion of each company. Life Technologies constructed a 1.5Kb mate pair library and Illumina constructed 2Kb, 2.5Kb, 3Kb and 4Kb libraries for sequencing. The SOLiD instrument provided 42.22 G bases of sequence corresponding to ~180X mean depth and the GAI gave 14.17 G bases for a total of ~76X mean depth.

SOLiD mate-pair library construction



GAI mate-pair library construction



Alignment of Reads

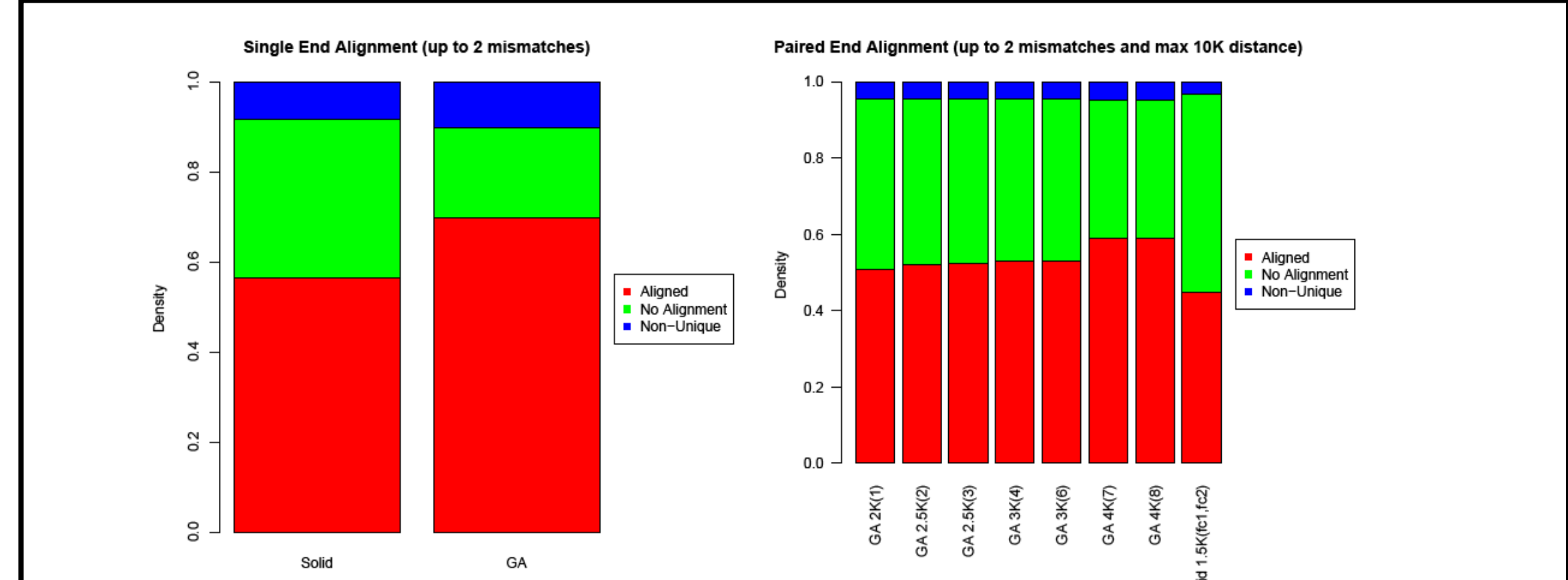


Figure 1. Reads were aligned using bowtie. Alignments were required to be unique, allowing up to two mismatches per read and a 10Kb distance between pairs.

Methods

End Sequence Profiling (ESP):
Many different types of mapping signatures may be encountered when analyzing ESP data. The most basic scenarios of simple insertions, deletions and concordance are depicted in figures 1-3 below.

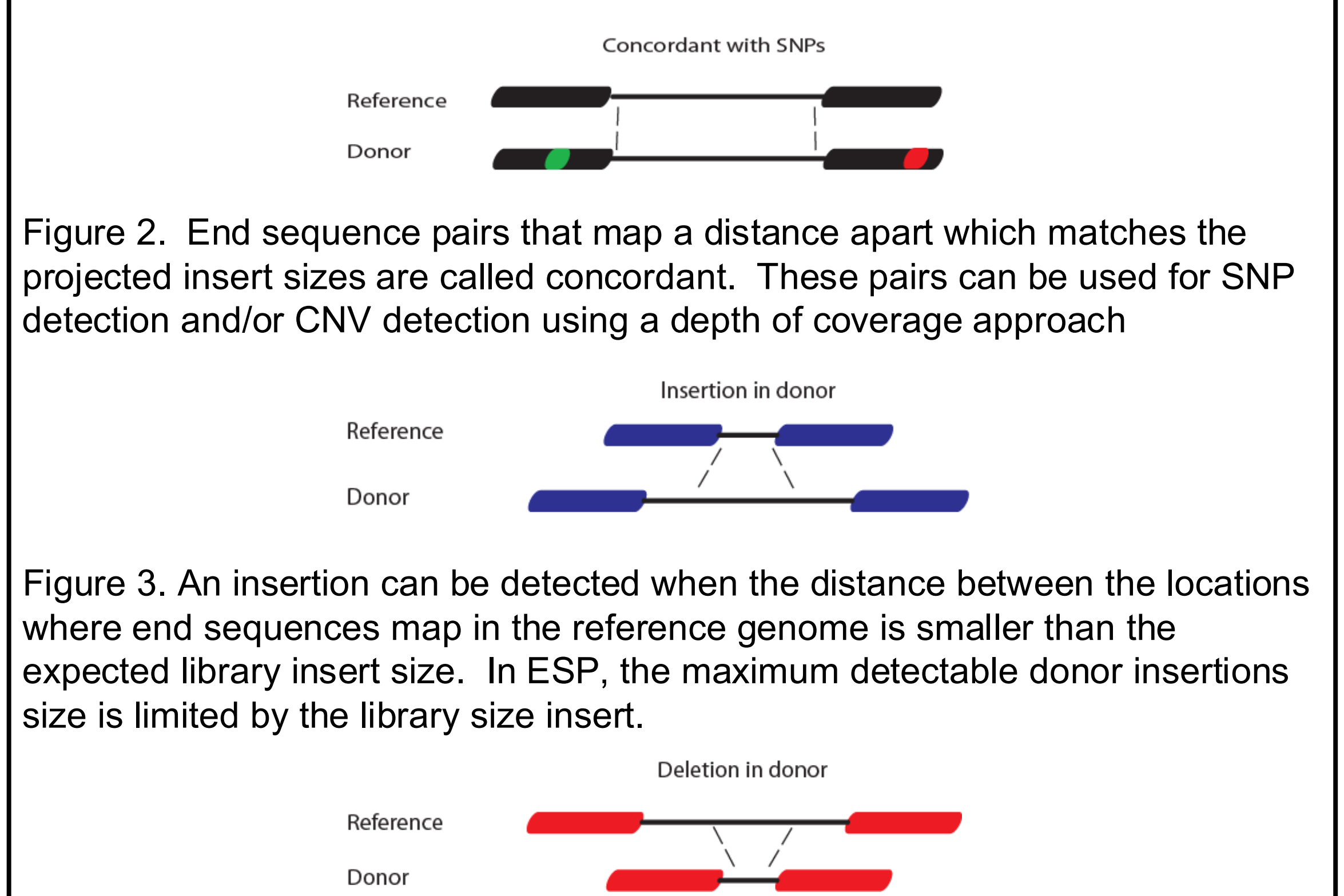


Figure 2. End sequence pairs that map a distance apart which matches the projected insert sizes are called concordant. These pairs can be used for SNP detection and/or CNV detection using a depth of coverage approach

Depth of Coverage (DOC):
HTS allows for the detection of gain/loss events by examining the levels of aligned reads throughout the genome. Both paired and un-paired reads can be used for detecting CNVs with the DOC measure.



Figure 5. A potential duplication signature where the number of reads mapped to a given location is much greater than expected given the total number of reads and genome size



Figure 6. A potential deletion signature where the number of reads mapped to a given location is zero or much lower than expected given the total number of reads and genome size

Split Mapping:
Indels can be contained internally within reads from a donor genome where the beginning and end of a read map to different areas of the genome. Due to the short reads of HTS, many false positive split mappings may occur. By assuming one end of a pair must map a given distance from the other, the false positive split mappings can be reduced. In addition, clusters of anchored split mappings can be used to increase the confidence of a prediction.

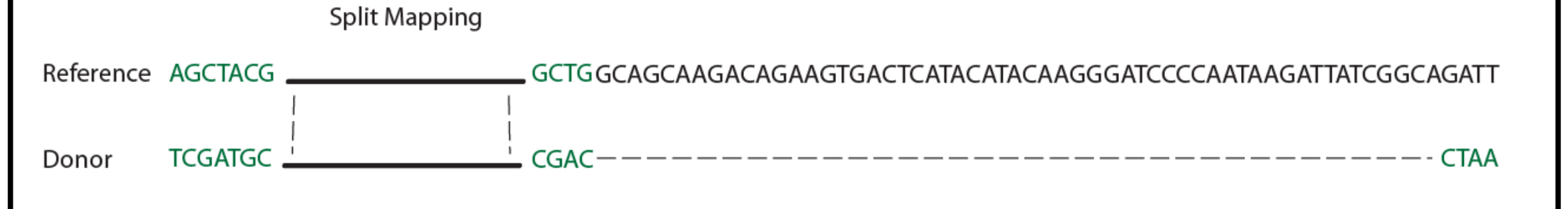
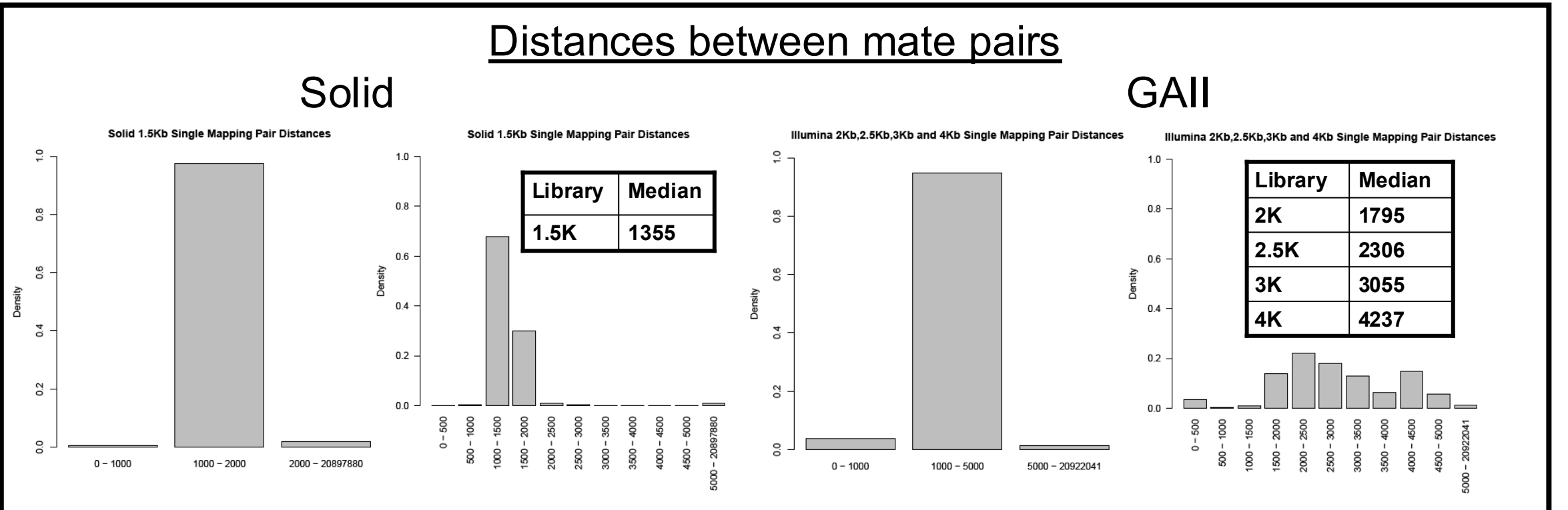


Figure 7. A deletion can be located within one end of a read pair by leveraging an alignment for the other end.

Results



CNVs found by Solid and GAI using ESP compared to aCGH



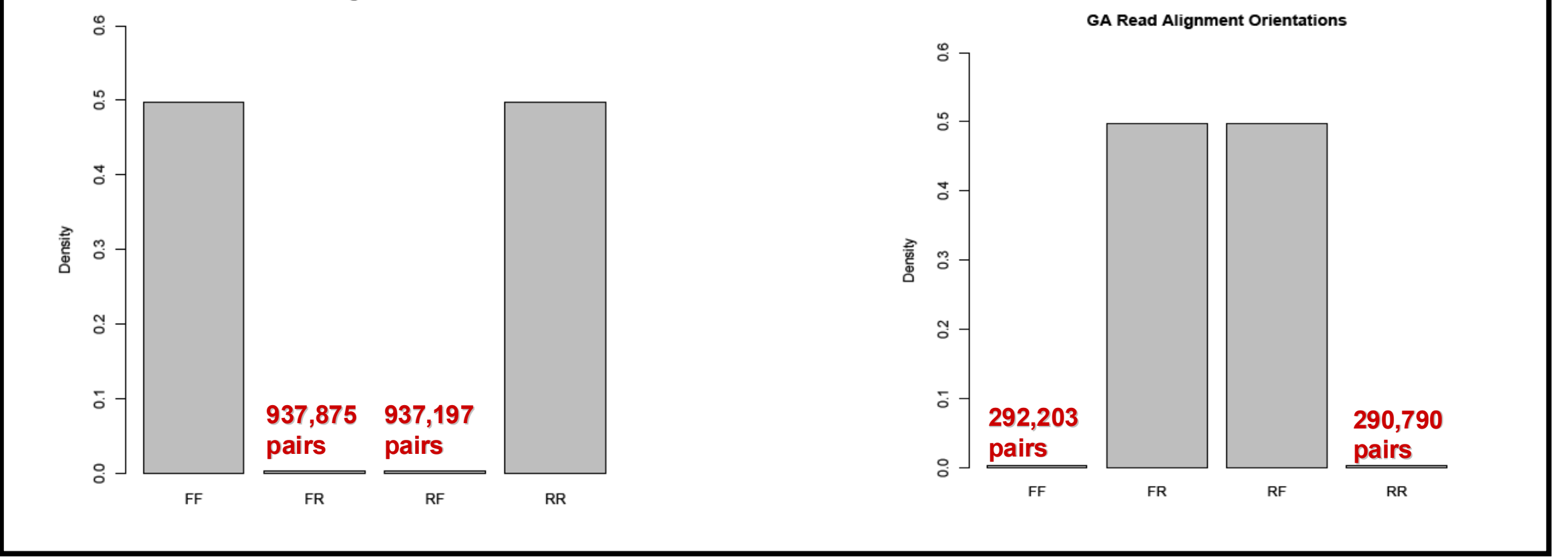
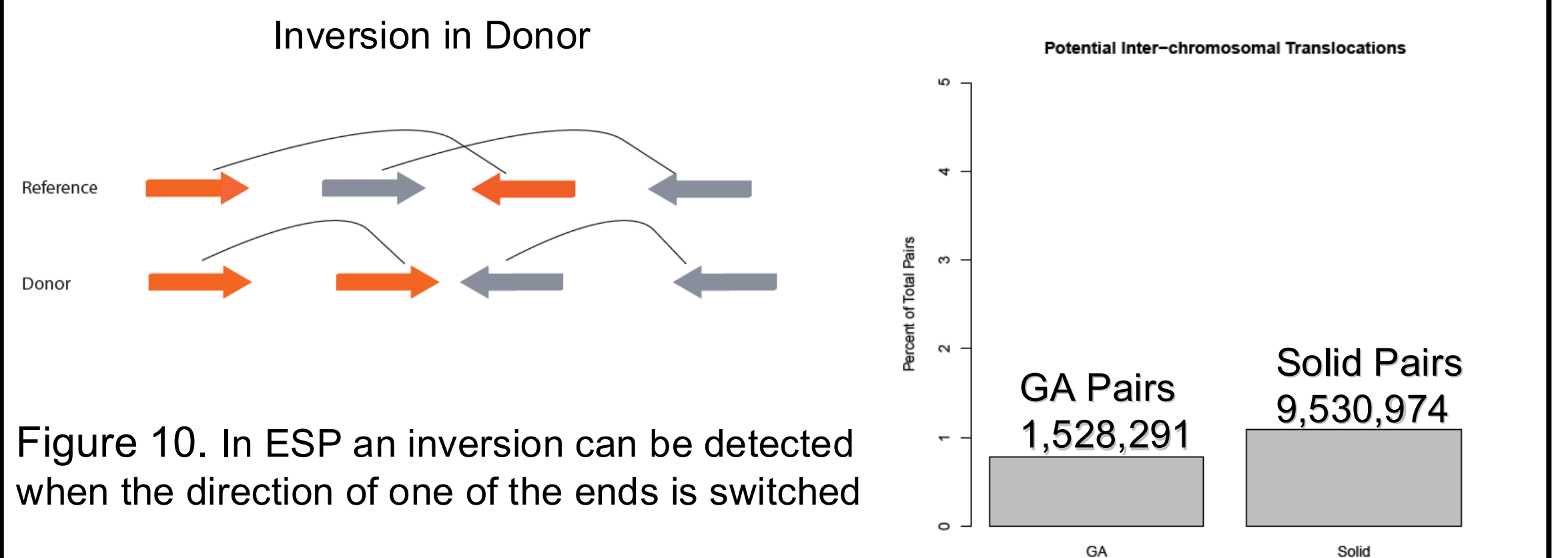
Figure 8. Using Breakdancer on the Solid and GA data, 409 of 531 Maydan CNVs were found

Future Directions - CNV Validation

Upon completion of our primary analysis, we will pick a number of CNVs to validate. These CNVs will consist of a subset of true positives (known CNV region detected by sequencing), false negatives (known CNV region not detected by sequencing), and false positives or novel CNVs (new CNV regions detected by sequencing). We will utilize LifeTechnology's CopyCaller workflow to validate the CNVs detected. This technique uses qPCR along with TaqMan® Copy Number assays to quantify the number of copies of a specific region. Both the Hawaiian and Bristol strains will be assessed for the CNVs of interest during the validation procedure.

Future Directions - Structural Variants

We will make use of our dataset to look at structural variants including inversions, translocations, and other genomic permutations.



We would like to thank

- Charles Cochran and Life Technologies for generating SOLiD data, expert data analysis and CNV validation using their CopyCaller workflow.
- Joel Fellis and Illumina for generation of Illumina GAI data.
- The HoYi Mak lab at the Stowers Institute for providing the *C. elegans* strains
- Ken Chen et al. for Breakdancer