

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

2009 Genomic Variation Research Group  
project




## Today's presentation

- Brian Sanderson – Intro to 2009 project
- Sean Blake – Methods used in study
- Aaron Noll - Data analysis performed



## Genomic Variation Research Group

- Bruce Kingham – 2009 Chair – U. of Delaware
- Sean Blake – U. of Missouri
- Casey Dagnall - NIH/NCI
- Helaman Escobar – MWG/Operon
- Amy Hutchinson – NIH/NCI
- Karen Jonscher – UC Denver
- Chris Lytle – Dartmouth
- Aaron Noll – Stowers Institute
- Brian Sanderson – Stowers Institute



## Thank You to our sponsors!

- Life Technologies
  - Mate pair library construction (1.5Kb) and NGS on SOLiD
  - CNV validation using TaqMan CNV assays and CopyCaller workflow
- Illumina
  - Mate-pair library construction (2, 2.5, 3 and 4Kb) and NGS on Genome Analyzer II
- HoYi Mak lab at Stowers Institute
  - Hawaii and Bristol strains of *C. elegans*



## 2009 GVRG study

- How is next generation sequencing filling the need for Copy Number Variant (CNV) detection and discovery?



## Why CNVs?

- Hot topic...with many important biological implications.
- Prevalence - estimated 0.4% of unrelated human genomes differ with respect to copy number (Redon et al, 2006)
- Cancer
  - CNVs have been shown to be present in many cancers
  - Have been used to predict metastatic capabilities of some cancers
  - Implicated in adaptive evolution processes such as resistance to chemotherapy (Engelman et al, 2007)
- Associated with numerous additional diseases
  - autism, schizophrenia, Alzheimers, Parkinsons, early onset obesity and susceptibility to HIV



- Definitions

- Aneuploid is a genome with extra or missing chromosomes
- CNV is “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.”
- InDel is “a collective abbreviation to describe relative gain or loss of a segment of one or more nucleotides in a genomic sequence...typically used to denote relatively small-scale variants”

S.W. Scherer, et.al. (2007). Challenges and standards in integrating surveys of structural variation. *Nat.Gen.* 39: S7-S15.



## How well does NGS fill the need for CNV discovery / detection?

- Hypothesis: an optimal combination of mate-pair gap and analysis methods will have the most traction in next-generation sequencing for CNV detection
- Sequence Hawaii strain of *C.elegans* and compare to published Bristol (canonical) strain
  - Will different sizes of mate pair gaps affect the comparison?
  - What coverage is necessary to discern 1Kb CNVs?
  - What tools are available for aligning and analyzing sequenced mate-pairs?
- Provide (from our viewpoint) a current snapshot of next-gen technology and analysis methods for CNV studies



## Basis for study

### Methods

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

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

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

17:337–347 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Genome Research 337  
www.genome.org

- Shows a ~ 2% difference in Hawaii vs Bristol strains of *C.elegans* due to “coding” CNVs
- Report 531 CNVs and give genomic location for each
- Gives a good starting point and allows us to compare our NGS results to their aCGH.



## Sean Blake



## Preparing to Sequence *C. elegans*

- *C. elegans* has 6 chromosomes.
- As a model system, *C. elegans* has been extensively studied and multiple strains fully sequenced/annotated.
- The genome has been well characterized. This allows for a good experimental assessment of CNV discovery using NGS.
- The genome is relatively small at ~100Mbp, making it cost effective to sequence.
- High amounts of CNV are detected at coverage levels which can be achieved with only a partial run on the GA or SOLiD platforms.
- An isogenic population of the *C. elegans* Hawaii strain was grown on *e. coli*.
- Genomic DNA was prepared from the population using a phenol/chloroform-based extraction technique adapted from Current Protocols in Molecular Biology (2003, Unit 13.11). Completed at the Stowers Institute.



## Sequencing Methods

- To test our hypothesis, we requested sequencing support from vendors with established high throughput sequencing platforms.
- Illumina (Genome Analyzer) and Life Technologies (SOLiD) graciously supported our suggested approach as follows.
- Mate Pair Library Preparation: Perform preps in-house, insert size was left to the discretion of the participating vendors
- Sequencing: Perform using the current platforms, amount of sequence generated was left to vendors discretion.
- Data Analysis: use approach that allows for assessing capability of HTS for CNV detection and comparison to existing aCGH data (for Hawaii strain)
  - Provide CNV analysis using tools that were either currently available or that could be described herein, in order to get a clear snapshot of tools available for such an application.
  - Use the Bristol reference genome for sequence alignments and detection of CNVs.
  - Summarize and compare the CNVs detected by sequencing to those discovered by Maydan et al (using aCGH technique).



## Validation Methods

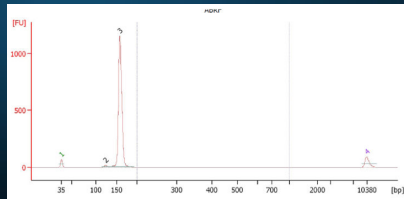
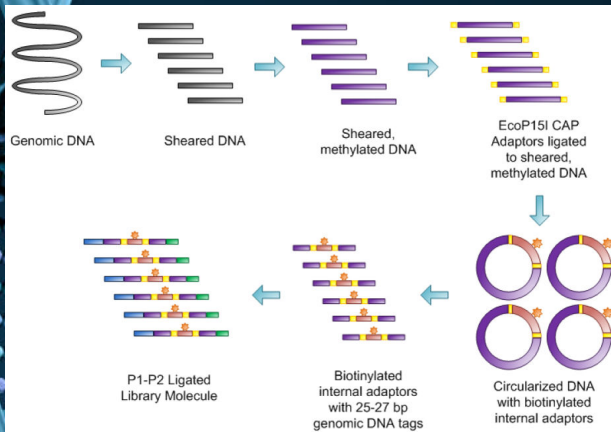
- In addition to sequencing, we requested assistance from Life Technologies in performing CNV validation.
- A subset of CNVs will be chosen to validate, consisting of true positives (known CNV region detected by sequencing), false negatives (known CNV region not detected by sequencing), and false positives 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 (to be run at NIH/NCI).
- Life technologies agreed to adapt the design process specifically include *C.elegans*.



## Mate Pair Library Generation

- Mate pair libraries were generated using existing protocols and kits.
- Sample preparation designed to enhance the # of short reads that map to the genome by constructing read pairs with known gap sizes. The large gap size and paired reads approach is specifically designed to produce signatures in the data analysis (ie alignment process) which ideally span CNVs and perhaps structural variations.
- M.P. library preparations are slightly more technically challenging sample preparation vs std small insert sequencing library (no automation for this prep).
- Multiple enzymatic, shearing and enrichment steps are unique to each vendor's method, and some steps are sensitive to template amounts (requiring an Agilent Bioanalyzer).

# SOLiD™ – Eco Mate Pair Library Construction Overview

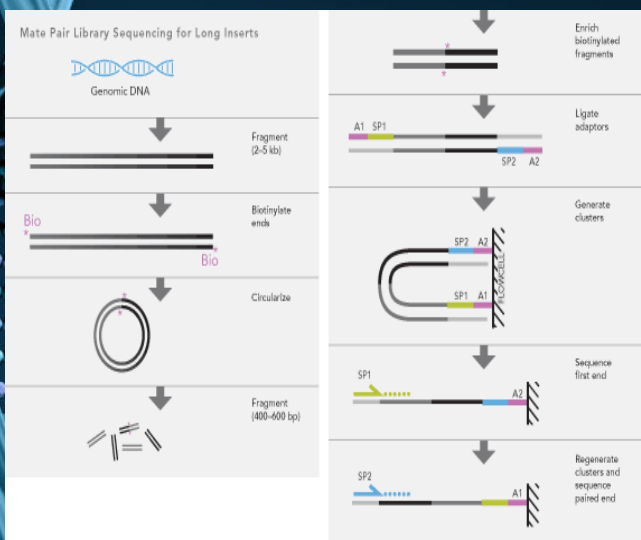


## Protocol Comments:

- 36 µg input DNA sheared to 1.5 kb-
- 1.9 µg remaining after size-selection. And used for circularization.
- 0.378 µg remaining after circularization.
- Adaptors (P1 and P2) are added to the sheared DNA for use with SOLiD™ system.
- 12 cycles of amplification applied to library.
- Bioanalyzer QC shows correct library size (156 bp).
- Greater than expected DNA loss during library construction.

Methods (cont'd)


# Illumina Genome Analyzer: 2-5kb Mate Pair Library Construction Overview



## Protocol Comments:

- Various mate pair libraries generated
  - 2, 2.5, 3 & 4kb
- 10 ug input DNA sheared to appropriate size using Hydroshear (nebulization or Covaris shearing can be used if optimized).
- End repair, biotin labeled nucleotide addition, and size selection are performed prior to circularization.
- Primary goal is to maintain library diversity by recovering ~200ng of material for the circularization step (using Bioanalyzer), which is sensitive to DNA concentrations that alter levels of chimeras and false mate pairs.
- Adaptors are added to biotin enriched fraction for cluster generation and sequencing on the Genome Analyzer IIx.
- Final library QC is recommended using Bioanalyzer (validation of correct library size and quantitation).
- Mate pair libraries are sequenced using the Paired End sequencing procedure

Methods (cont'd)



## Mate Pair Sequencing and Library Diversity

- High quality sequence information was generated using:
    - **ABI SOLiD 3 Plus: 1.5kb M.P. library, 2x25bp, one full run (2 Flowcells)**
    - **Illumina Genome Analyzer Iix: 2,2.5,3 & 4kb M.P. libraries, 2x36bp paired end, one full run (1 Flowcell, 1lane 2kb, 2lanes 2.5,3,4kb)**
  - A 'good' library preparation would result in a high percentage of fragments in the final library which upon alignment have close to the expected gap sizes, and which have effectively avoided large % of duplicate reads during the prep ensuring a high diversity of fragments (ie unique reads) from the genome in the proper orientation.
  - Each sequencer will generate reads from the sample prep which reflect the amounts of true and false mate pairs generated from the prep based on the expected alignment orientation.
  - The experimental data can be used to calculate levels of library diversity by taking into consideration:
    - SOLiD: # M.P., median gap size, # normal M.P., # non-redundant normal M.P.
    - Illumina: Median gap size, % variation, % innies (false m.p.), % outies (true m.p.), % duplicates, % chimeras (smallest 2kb M.P. sizes show the best statistics for all)
- Methods (cont'd)



## Aaron Noll



## Agenda

- Identify specific analysis objectives
- Discuss possible CNV detection approaches
- Analysis methods
- Observations
- Preliminary results
- Conclusions
- Future Directions




## Data Analysis Objectives

- Analyze Solid and GA datasets using common set of tools
- Investigate how read quantity affects CNV detection
- Identify novel deletions for validation



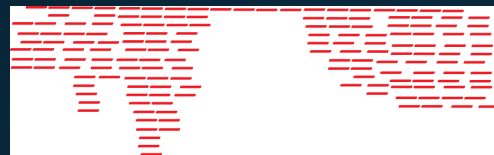
## Common CNV Data Analysis Approaches

- DOC – Depth of Coverage
  - Similar to aCGH
- ESP – End Sequence Profiling
  - Relies on read pairs
- Split read mapping
  - Scan for indels within reads

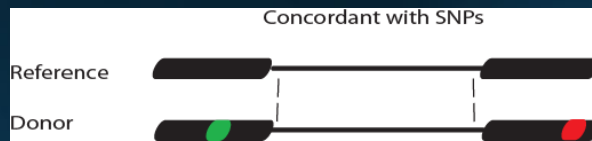


## DOC – Depth of Coverage

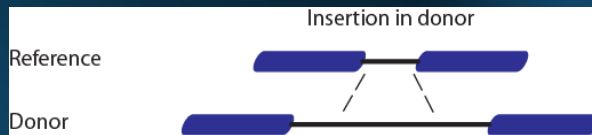
- Like in aCGH, two samples can be compared to detect relative gains and losses of genetic material.
- In the case of a single donor, one may assume a uniform sequencing process, where the number of reads mapping to a region follows a specific distribution and is expected to be proportional to the number of times the region appears in the donor



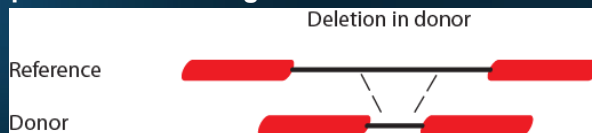
# ESP - End Sequence Profiling



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



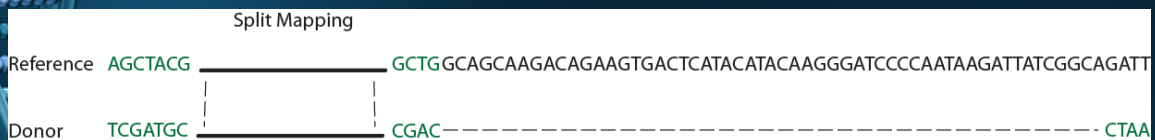
An insertion can be detected when the distance between the locations where end sequences map in the reference genome is smaller than the expected insert size



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

# Split Read 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
- Because reads are short, many false positive split mappings may occur. By assuming one end of a pair must map a given distance from the other however, the false positive split mappings can be reduced significantly





## Examples of different applications

- DOC only
  - Chiang et al. 2008 (sequenced both reference and donor)
- ESP only
  - Korbelt et al. 2007
- Split read mapping only
  - Ley et al. 2008
- ESP and split read mapping
  - Clark et al. 2010
- ESP, DOC and split read mapping
  - McKernan et al. 2009



## Tools – open source examples

- DOC – Depth of Coverage
  - Solid CNV tool
  - CNVSeq (requires sequence for donor and reference)
- ESP – End Sequence Profiling
  - BreakDancerMax
  - PEMer
  - Variation Hunter
  - Solid Large Indel tool
- Split read mapping
  - Pindel
  - BreakdancerMini



## Tools – open source examples

- DOC
  - **Solid CNV tool**
    - CNVSeq (requires sequence for donor and reference)
- ESP
  - **BreakdancerMax**
    - PEMer
    - Variation Hunter
    - Solid Large Indel tool
- Split read mapping
  - Pindel
  - BreakdancerMini



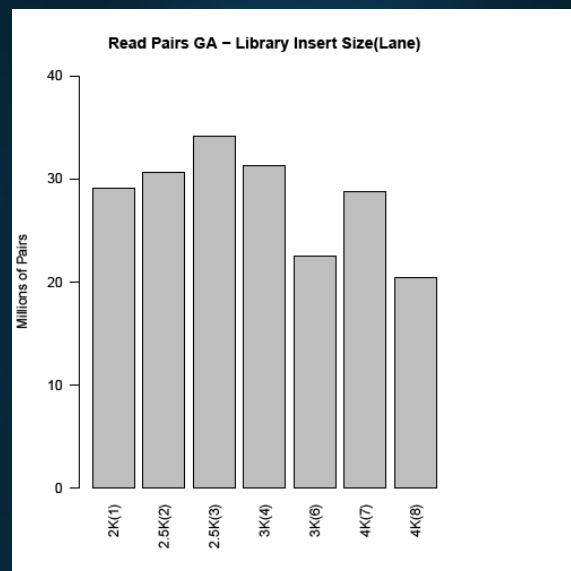
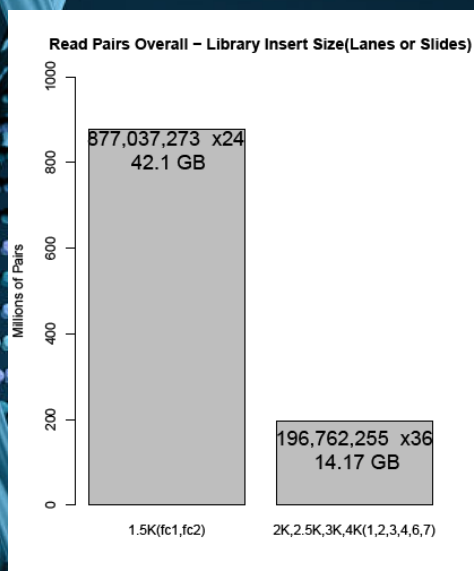
## Tools – commercial examples

- SoftGenetics
  - DOC

## Methods for deletion discovery

1. Align reads with bowtie allowing up to two mismatches per read and a maximum of 10K distance between pairs
2. Randomly select aligned reads at different quantities and analyze with BreakDancer (ESP)
3. Compare results to aCGH
4. Identify novel deletions found independently in Solid and GA datasets
5. Annotate results

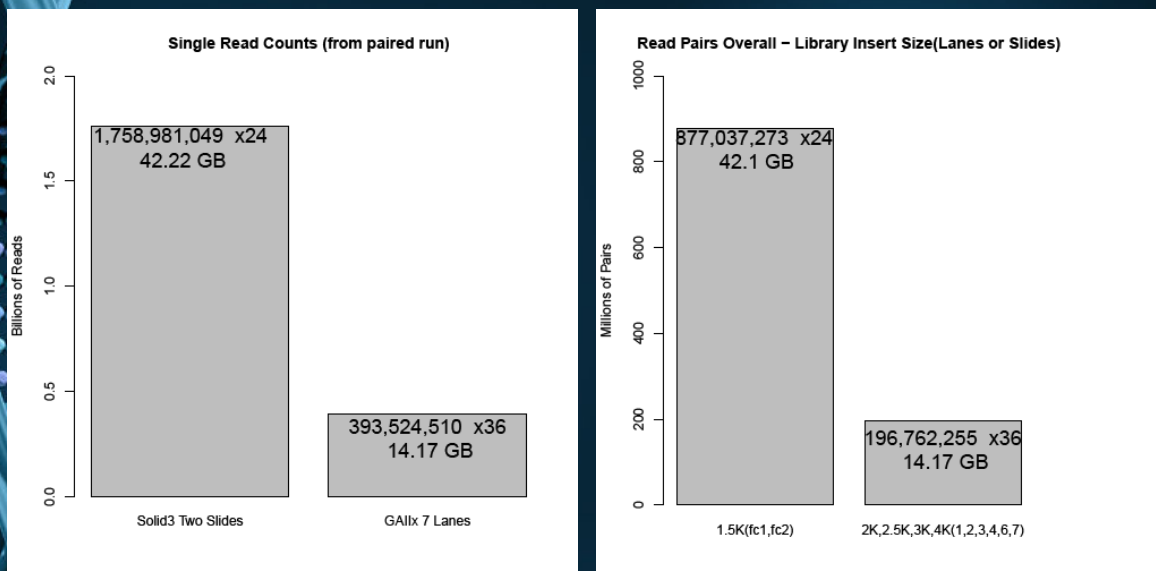
## Billions of bps from Solid and GA



# Salvaging Unpaired Reads

- In “PF” s\_N\_N\_sequence.txt files, Illumina only reports read pairs but no single end reads
- Life Technologies reports both paired and single reads
  - 1.6M and 3.3M single end reads were found for Solid fc1 and fc2 respectively
- If a DOC analysis approach were applied to the data, the extra single end reads from Life Technologies could be used

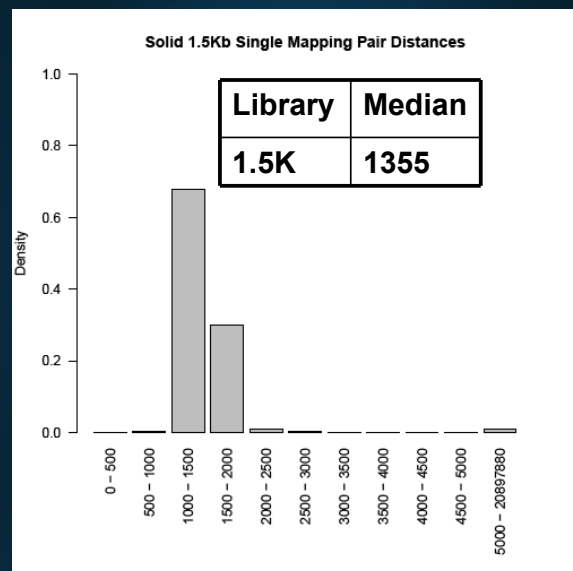
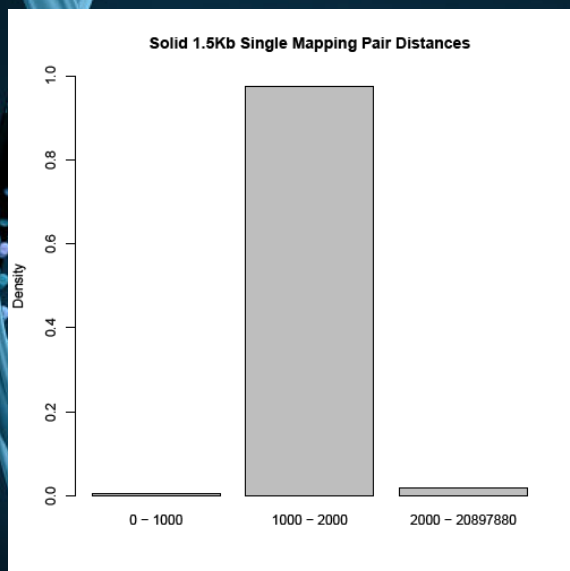
# Single Ends – Rescued ~ 0.1 GB



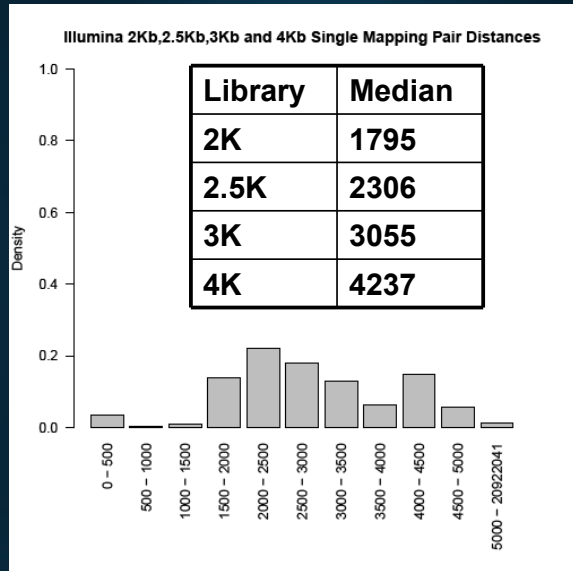
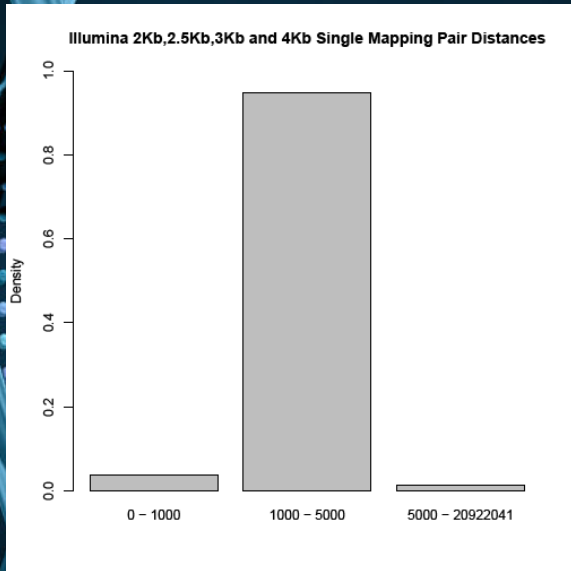
## Pair Orientation – depends on protocol

- GA Mate Pair
  - Reverse Forward <--->
- GA Paired End
  - Forward Reverse >--<
- Solid Mate Pair
  - Forward Forward >-->

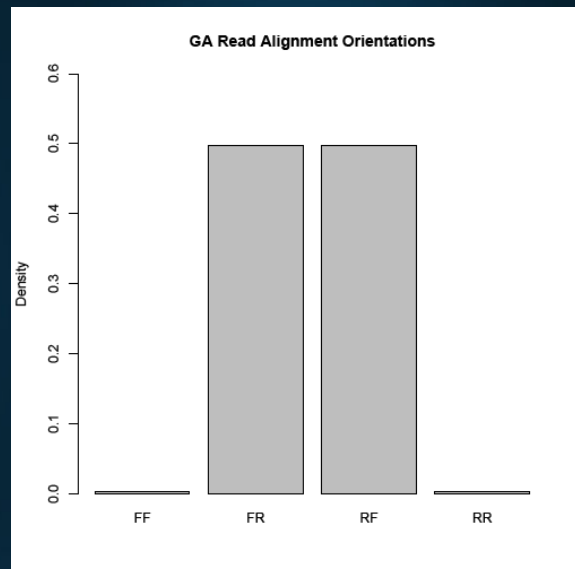
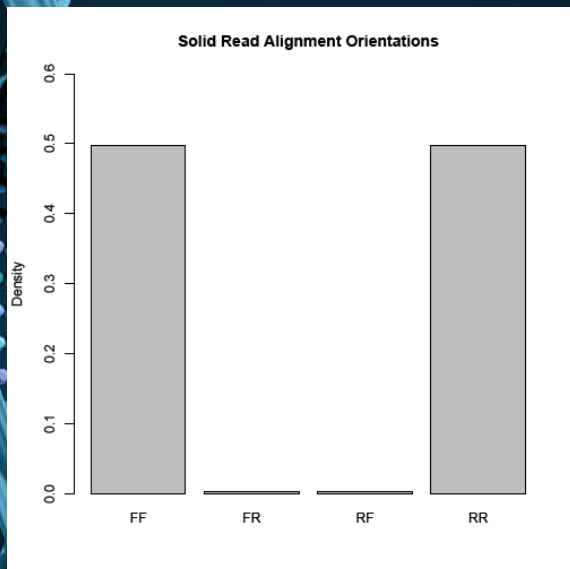
## Distances between pairs Solid – max 10K



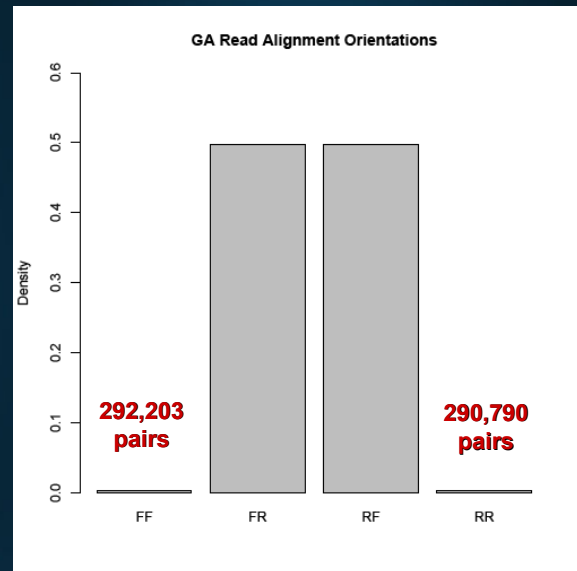
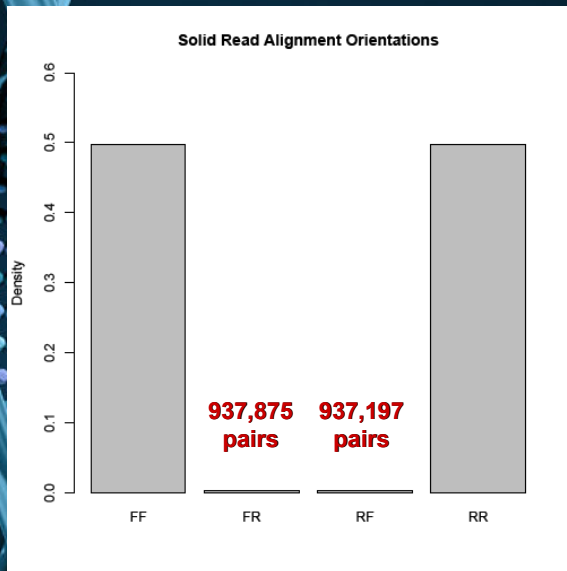
# Distances between pairs GA – max 10K



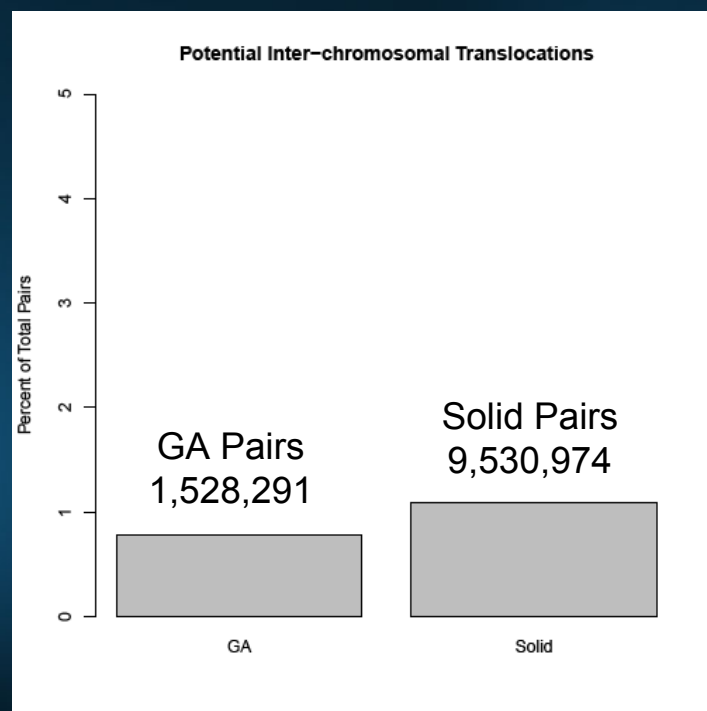
# Paired Orientations In Data



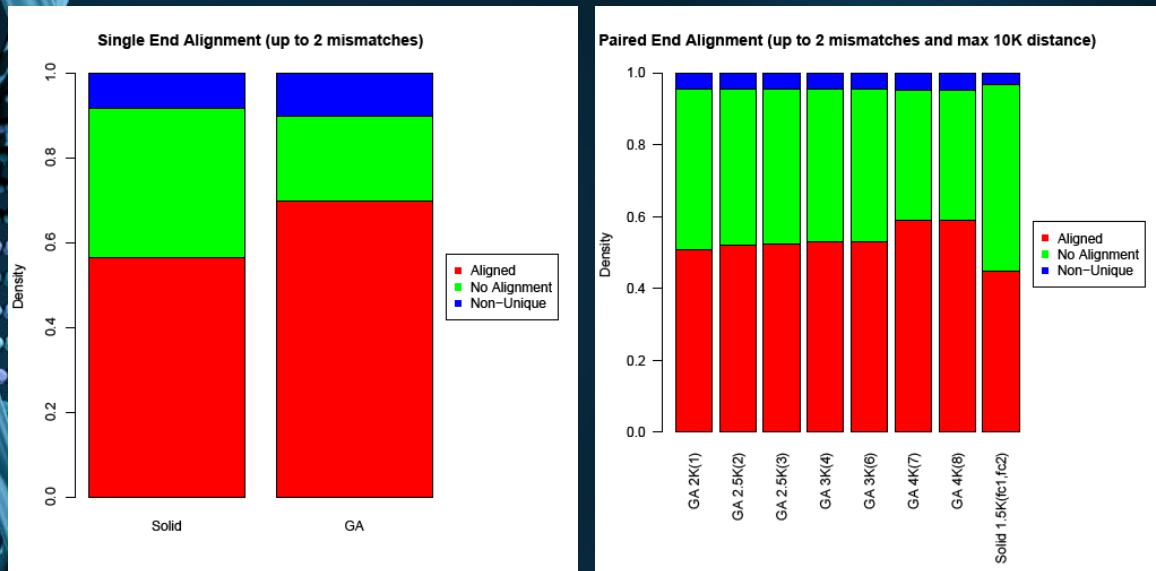
# Paired Orientations In Data – SVs?



# Potential Inter-chromosomal Translocations



# Alignment Results – single and paired



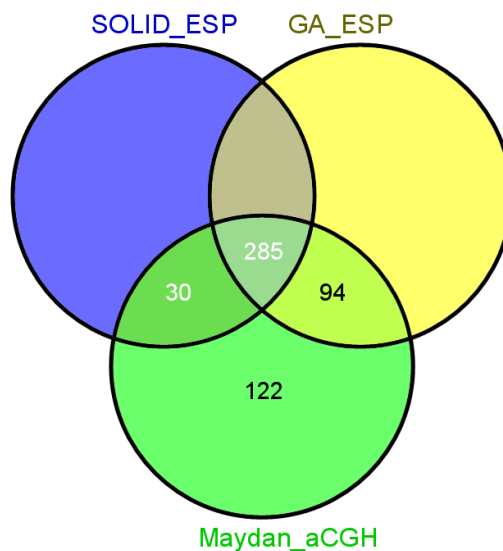
# Mean ~180X and ~76X depth

Solid Paired					
Chr	Length	Bases Covered	Mean Depth	% Covered	Total Sequence Data (bps)
I	15080552	14561797	196.0411757	96.56010602	2,956,409,144
II	15279311	14276304	183.640949	93.43552206	2,805,907,172
III	13783317	13264431	182.6187068	96.23540545	2,517,091,526
IV	17493785	16744095	180.3777566	95.71453519	3,155,489,692
V	20922231	19030668	173.9142453	90.96907602	3,638,674,014
X	17718850	17356972	170.0571986	97.95766655	3,013,217,994
GA Paired					
Chr	Length	Bases Covered	Mean Depth	% Covered	Total Sequence Data (bps)
I	15080552	14604361	87.79268398	96.84235033	1,323,962,136
II	15279311	14298517	76.92792011	93.58090165	1,175,405,616
III	13783317	13292468	75.56051479	96.43881803	1,041,474,528
IV	17493785	16823311	73.84604029	96.16735886	1,291,846,752
V	20922231	19083414	73.18599436	91.21118106	1,531,214,280
X	17718850	17398842	72.34049027	98.19396857	1,281,790,296

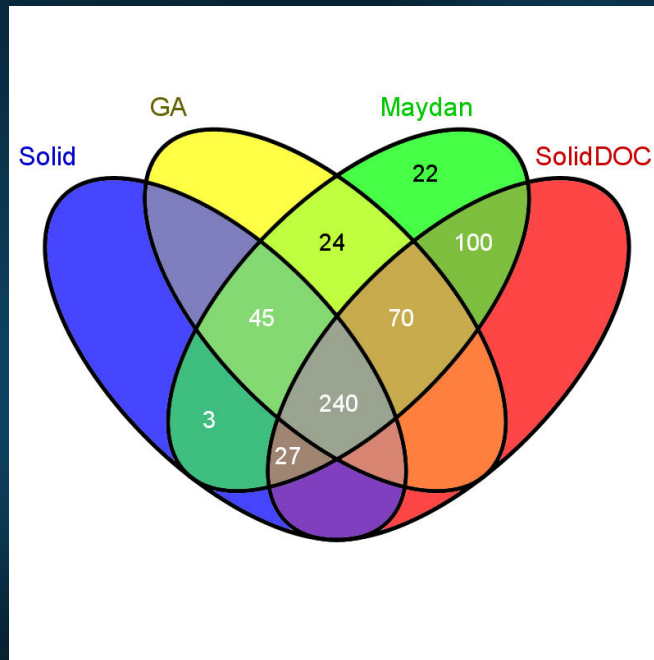
# Longer reads and insert sizes yield greater coverage with less data

Solid Paired					
Chr	Length	Bases Covered	Mean Depth	% Covered	Total Sequence Data (bps)
I	15080552	14561797	196.0411757	96.56010602	2,956,409,144
II	15279311	14276304	183.640949	93.43552206	2,805,907,172
III	13783317	13264431	182.6187068	96.23540545	2,517,091,526
IV	17493785	16744095	180.3777566	95.71453519	3,155,489,692
V	20922231	19030668	173.9142453	90.96907602	3,638,674,014
X	17718850	17356972	170.0571986	97.95766655	3,013,217,994
GA Paired					
Chr	Length	Bases Covered	Mean Depth	% Covered	Total Sequence Data (bps)
I	15080552	14604361	87.79268398	96.84235033	1,323,962,136
II	15279311	14298517	76.92792011	93.58090165	1,175,405,616
III	13783317	13292468	75.56051479	96.43881803	1,041,474,528
IV	17493785	16823311	73.84604029	96.16735886	1,291,846,752
V	20922231	19083414	73.18599436	91.21118106	1,531,214,280
X	17718850	17398842	72.34049027	98.19396857	1,281,790,296

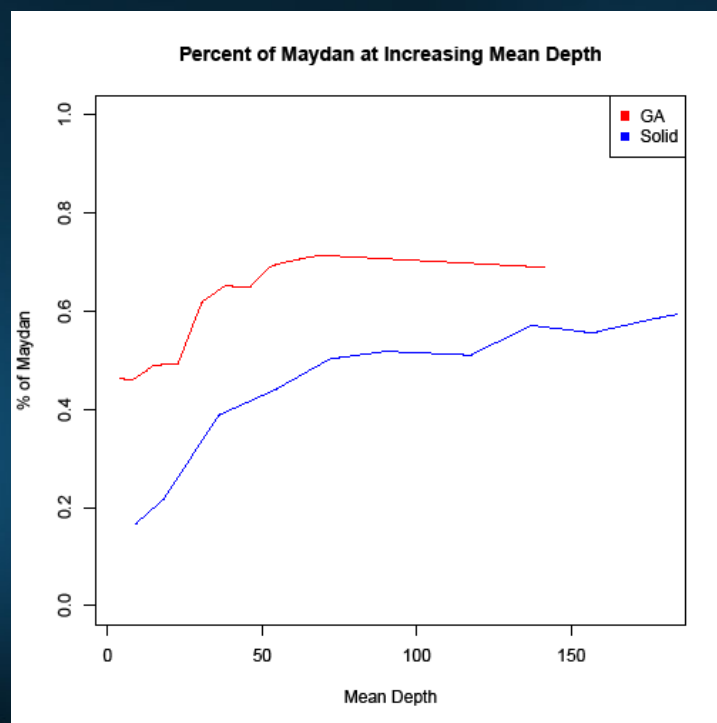
# aCGH deletions found by Solid/GA



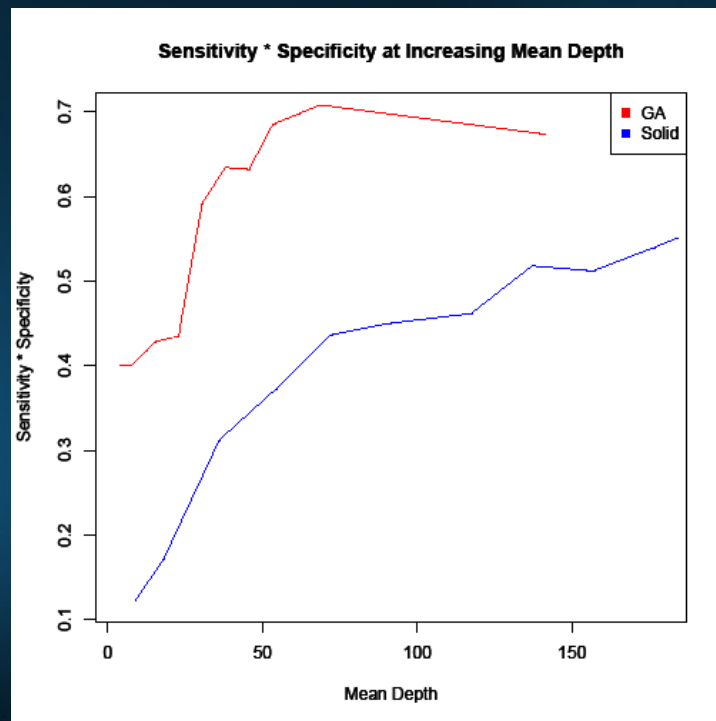
# 100 extra found by Solid CNV DOC tool



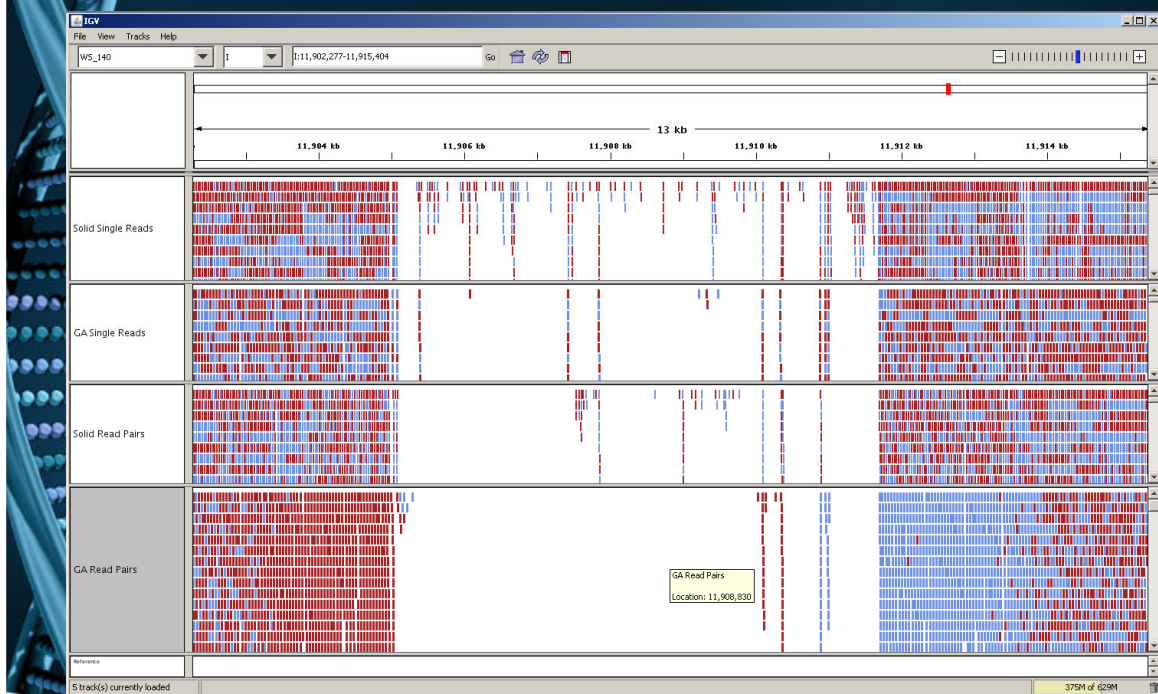
# ESP Results - TPs by mean depth



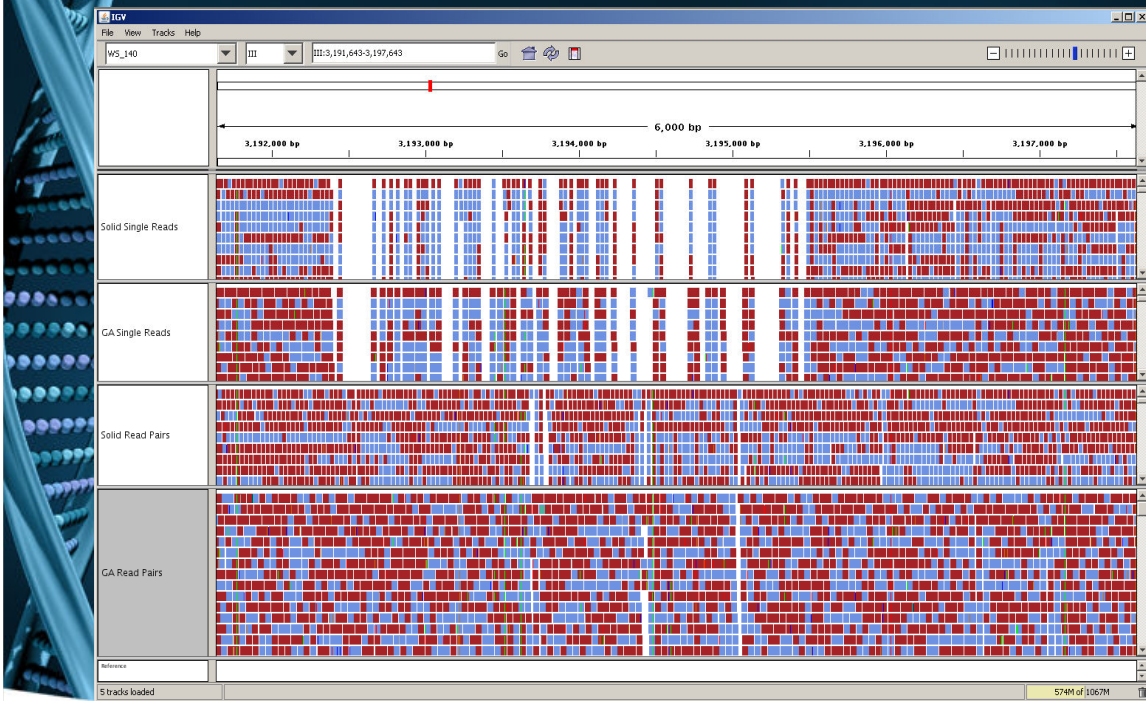
# ESP Results: Sens\*Spec by mean depth



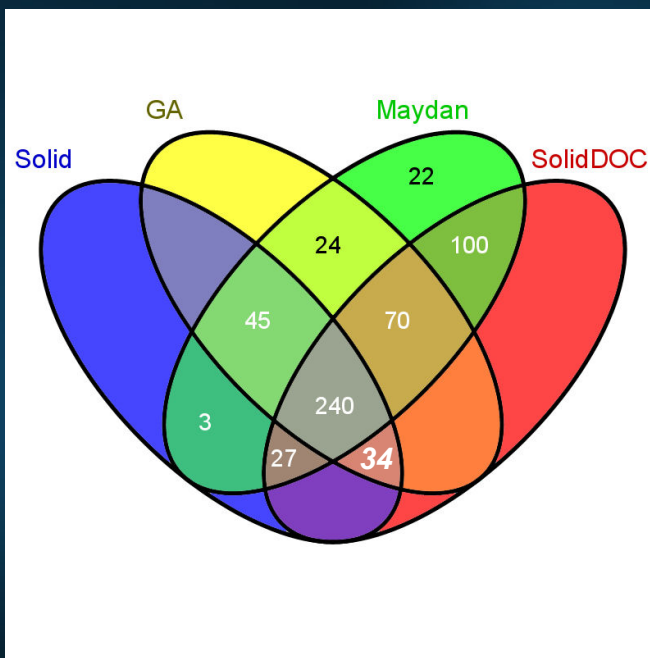
# Found by ESP but not Solid DOC



# Found by aCGH but not Solid or GA



# 34 novel deletions for validation



# Annotation - 34 novel putative deletions

region	size	coding genes	non-coding genes
I:1055615-1061413	5799	1	0
I:1206342-1216169	9828	2	0
I:1370772-1374887	4116	1	0
I:1754499-1763559	9061	1	0
I:2483606-2484154	549	2	0
I:3054255-3055132	878	1	0
I:3055910-3065872	9963	2	0
II:12597320-12602553	5234	1	0
II:13427947-13437782	9836	3	0
II:2122447-2127063	4617	0	1
II:2134037-2148748	14712	6	0
II:3580684-3584898	4215	3	0
III:12002419-12013806	11388	1	0
III:12065515-12074757	9243	2	0
III:2165649-2166173	525	1	0
III:2166423-2168225	1803	1	0
III:2351668-2361466	9799	1	0
III:2523885-2526421	2537	1	0
III:3078320-3086593	8274	1	0
IV:1348969-1352758	3790	1	0
IV:15856071-15860000	3930	0	19
IV:1601883-1602657	775	0	0
IV:16126218-16130753	4536	0	7
V:16980322-16985058	4737	1	1
V:17883851-17888506	4656	1	0
V:19674708-19684406	9699	1	0
V:2720746-2726707	5962	4	0
V:3773657-3779204	5548	3	0
X:1198179-1199682	1504	1	0
X:1395528-1400730	5203	2	0
X:1793690-1798865	5176	2	0
X:3562961-3567775	4815	2	0
X:4530912-4531662	751	0	0
X:5905985-5913284	7300	0	0

# Annotation - 34 novel putative deletions

coding - 28  
non-coding - 4

region	size	coding genes	non-coding genes
I:1055615-1061413	5799	1	0
I:1206342-1216169	9828	2	0
I:1370772-1374887	4116	1	0
I:1754499-1763559	9061	1	0
I:2483606-2484154	549	2	0
I:3054255-3055132	878	1	0
I:3055910-3065872	9963	2	0
II:12597320-12602553	5234	1	0
II:13427947-13437782	9836	3	0
II:2122447-2127063	4617	0	1
II:2134037-2148748	14712	6	0
II:3580684-3584898	4215	3	0
III:12002419-12013806	11388	1	0
III:12065515-12074757	9243	2	0
III:2165649-2166173	525	1	0
III:2166423-2168225	1803	1	0
III:2351668-2361466	9799	1	0
III:2523885-2526421	2537	1	0
III:3078320-3086593	8274	1	0
IV:1348969-1352758	3790	1	0
IV:15856071-15860000	3930	0	19
IV:1601883-1602657	775	0	0
IV:16126218-16130753	4536	0	7
V:16980322-16985058	4737	1	1
V:17883851-17888506	4656	1	0
V:19674708-19684406	9699	1	0
V:2720746-2726707	5962	4	0
V:3773657-3779204	5548	3	0
X:1198179-1199682	1504	1	0
X:1395528-1400730	5203	2	0
X:1793690-1798865	5176	2	0
X:3562961-3567775	4815	2	0
X:4530912-4531662	751	0	0
X:5905985-5913284	7300	0	0



## Conclusions

- 30X depth appears to provide the best cost/discovery ratio for ESP
- Longer reads and insert sizes allow for greater coverage at less depth
- 46% of aCGH deletions were found in GA data at ~ 3X depth
- Found all but 22 (~96%) of aCGH deletions
- For thorough examination of data best to use the DOC, ESP and split read mapping



## Potential Future Directions

- More in depth mining of data for CNVs and SVs using ESP, DOC and Split Mapping
- Look for overlaps between results
- Annotate regions of variation
- Validate interesting predictions