Whole-Exome Sequencing: Technical Details

Jim Mullikin
Director, NIH Intramural Sequencing Center
Head, Comparative Genomics Unit

Whole Exome Sequencing, Why?

- Focuses on the part of the genome we understand best, the exons of genes
- Exomes are ideal to help us understand high-penetrance allelic variation and its relationship to phenotype.
- A whole exome is 1/6 the cost of whole genome and 1/15 the amount of data

Biesecker *et al*. *Genome Biology* 2011, **12**:128

Twinbrook Research Building

NISC occupies entire 5th floor

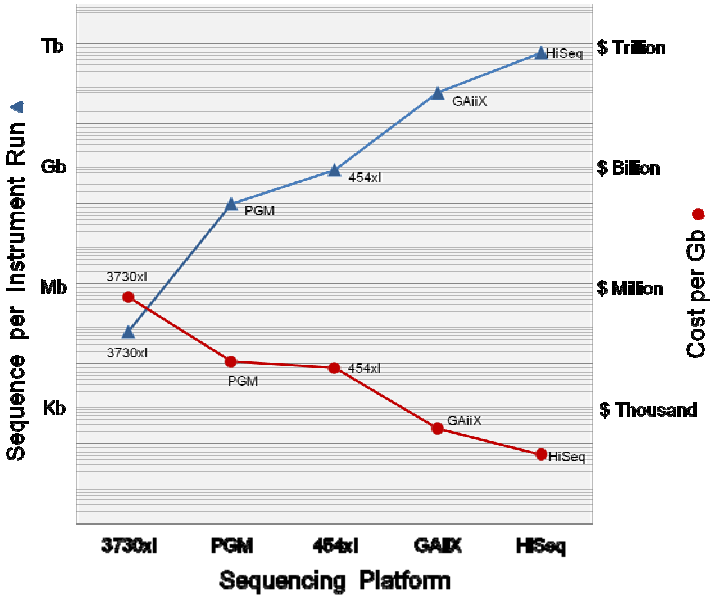
5625 Fishers Lane, Rockville MD

NISC Sequence Production

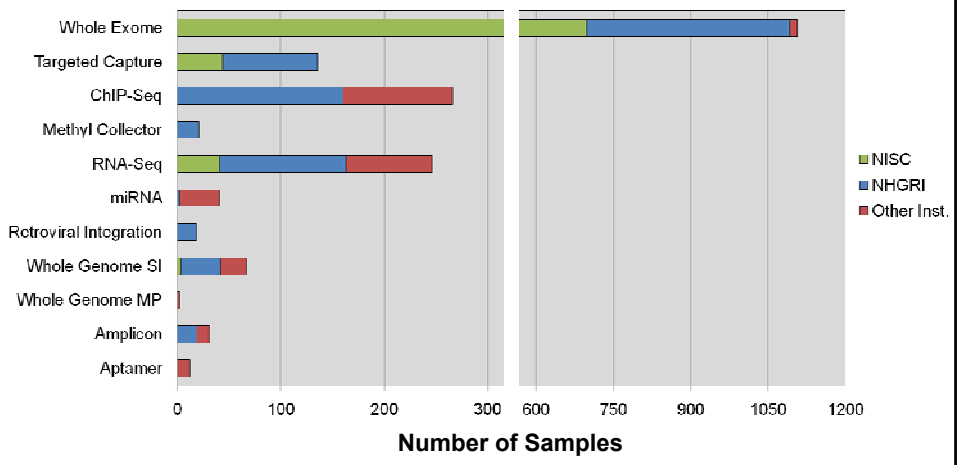
Feb. 2010

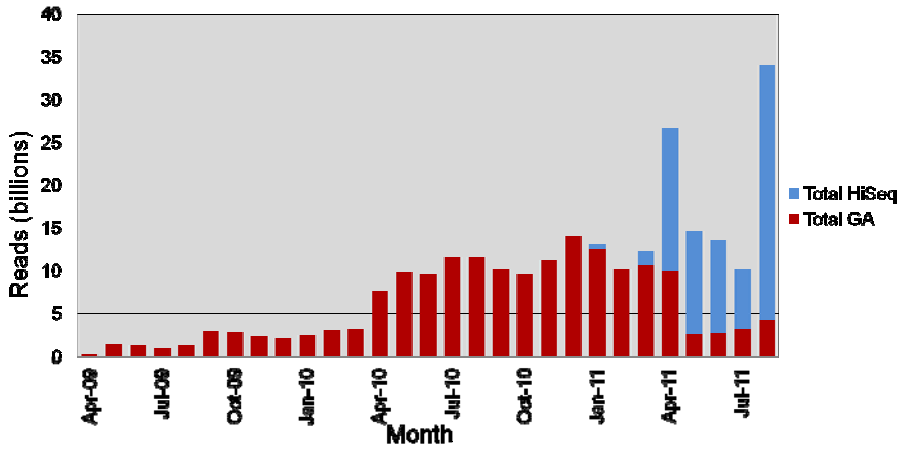
NISC Sequence Production

March 2011



Sequencing applications processed through NGS production pipeline from April 2009 to August 2011.

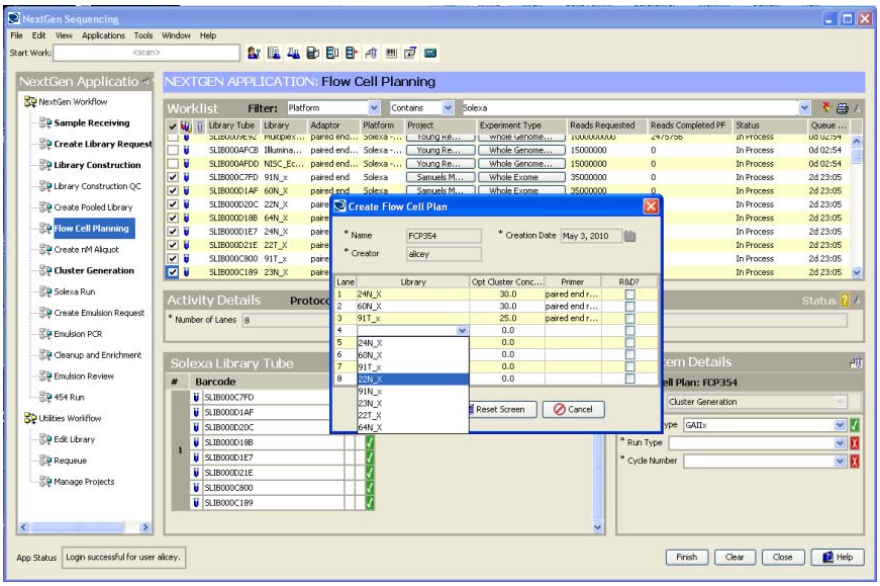




Total FY11 Sequenced : **169 billion reads**

Total NISC Lifetime Sequenced : **255 billion reads**

Cimarron Software based NextGen LIMS

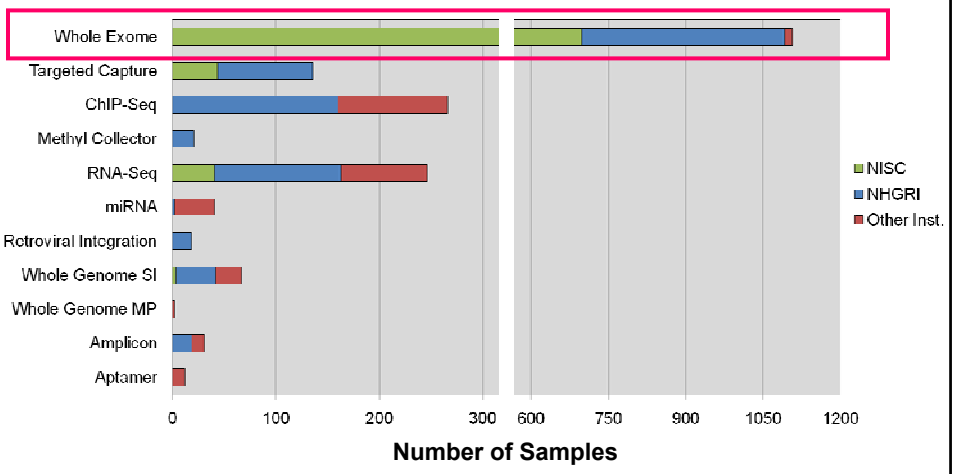


Computational Resources

for 6 GAiiX and 3 HiSeq2000

- **Linux cluster**
  - **1000 cores**
    - **250 for production**
  - **900TB disk**
    - **250TB for production with 75TB available**
    - **15TB/month long term storage**
  - **Network**
    - **1 and 10 Gigabit-Ethernet**

Sequencing applications processed through NGS production pipeline from April 2009 to August 2011.



Exome Sequencing Pipeline

Sample DNA Fragmentation

Illumina Library Preparation

Exome Enrichment

Cluster Generation

Sequencing and Basecalling

Sequence Read Alignment

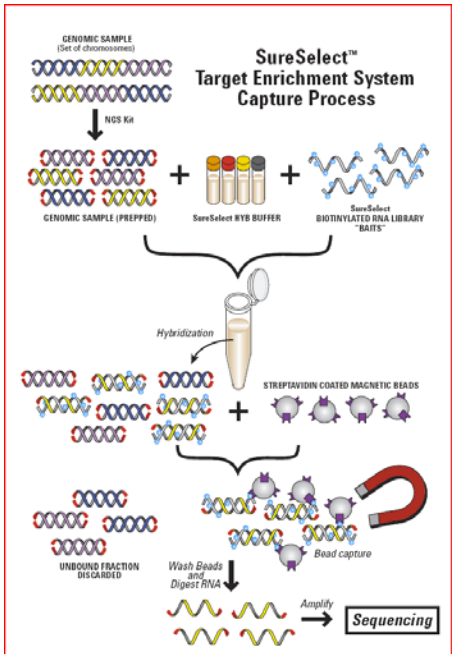
Variation Detection

Library Preparation

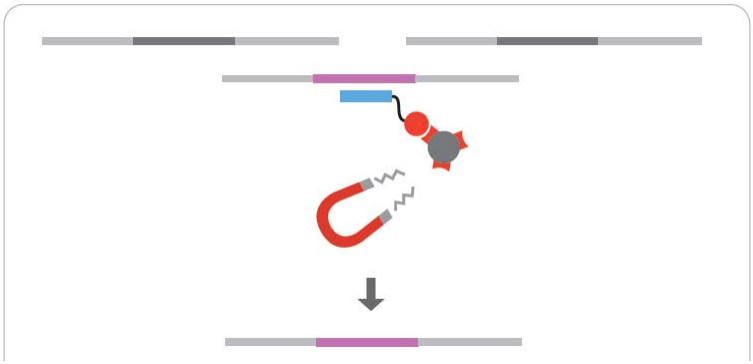
http://www.illumina.com/support/literature.ilmn

Agilent Technologies SureSelect method

Whole-exome kit 38Mb and 50Mb



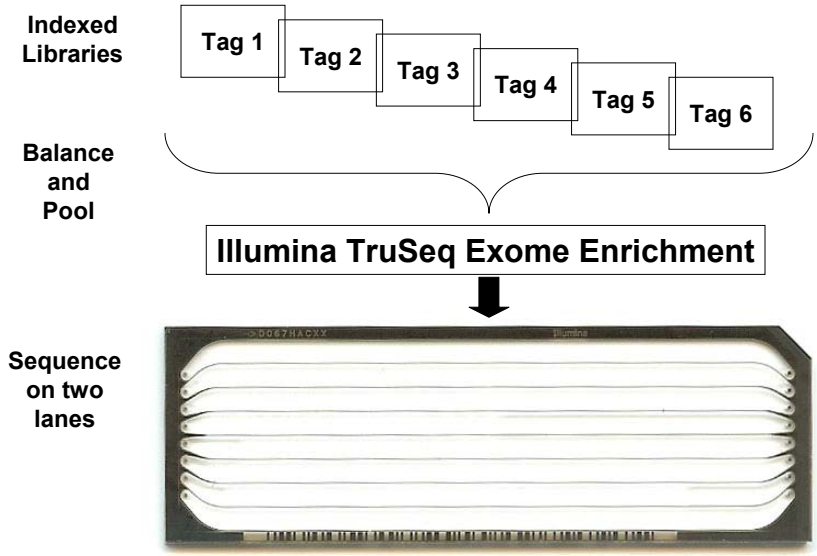
http://cp.literature.agilent.com/litweb/pdf/5990-3532EN.pdf



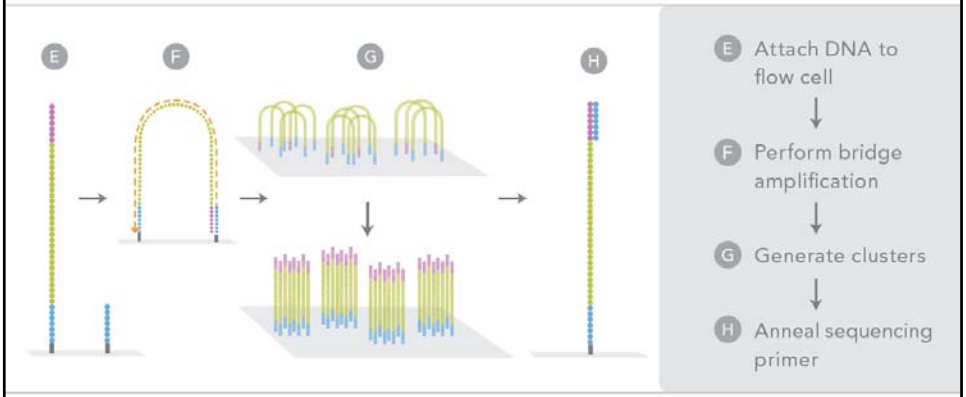
- Illumina TruSeq Exome Enrichment
- 62Mb of exome targeted

http://www.illumina.com/support/literature.ilmn

NISC Exome Process

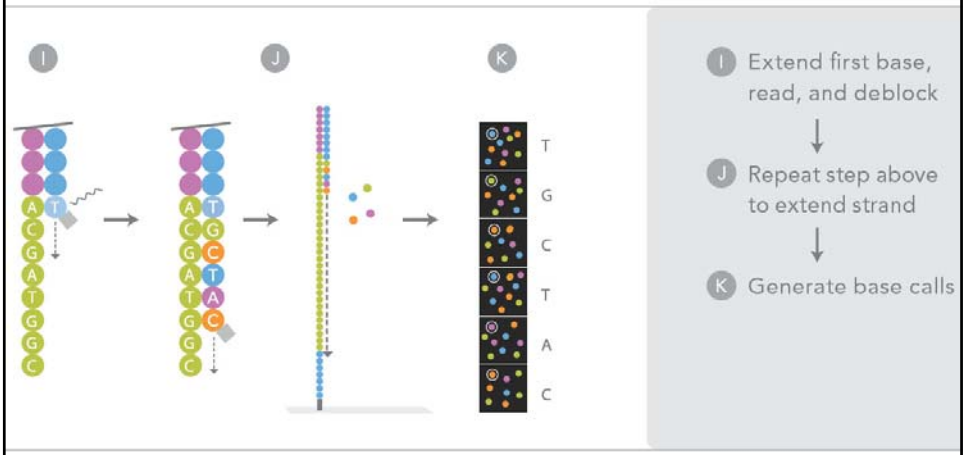


Cluster Generation



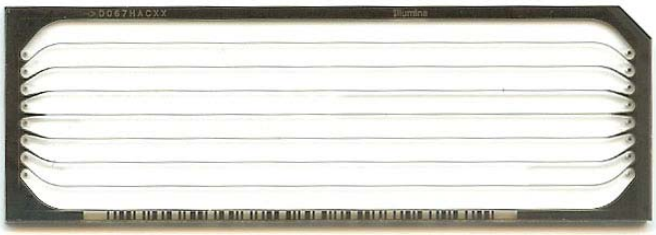
http://www.illumina.com/support/literature.ilmn

Sequencing



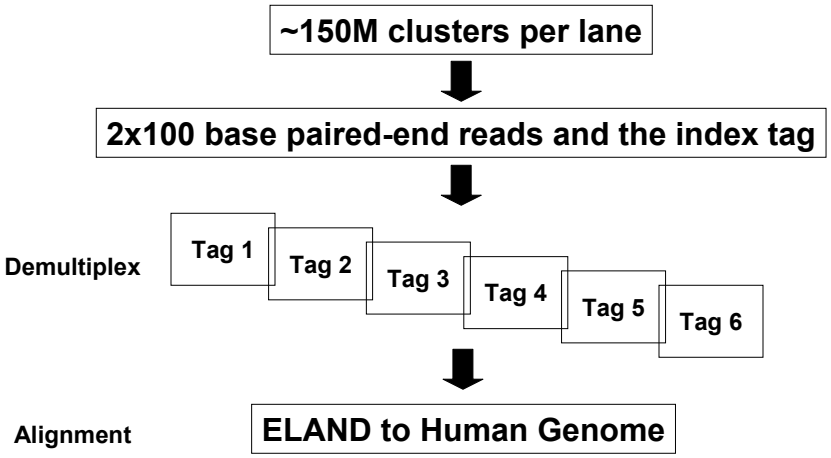
http://www.illumina.com/support/literature.ilmn

HiSeq Flow Cell



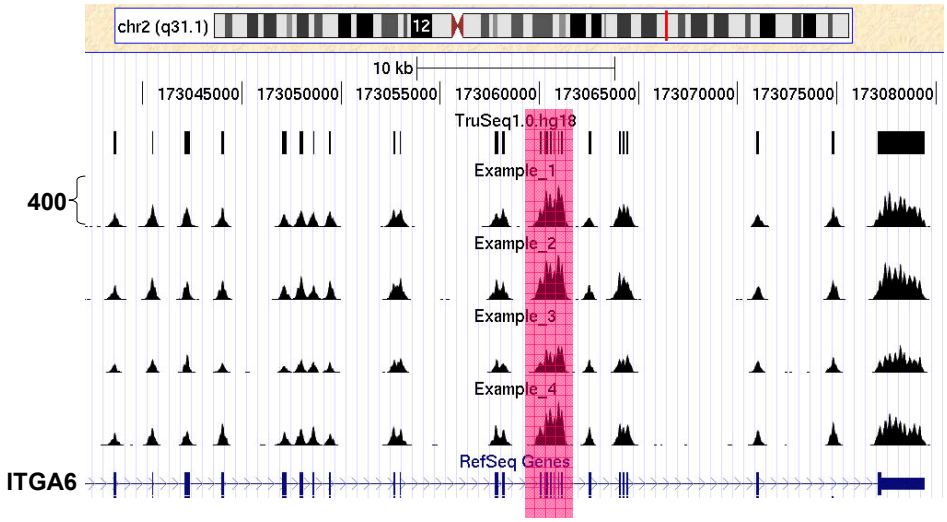
- 8 lanes
- 1.5 billion clusters
- Up to 300Gb per flow cell

Sequencing and Data Processing

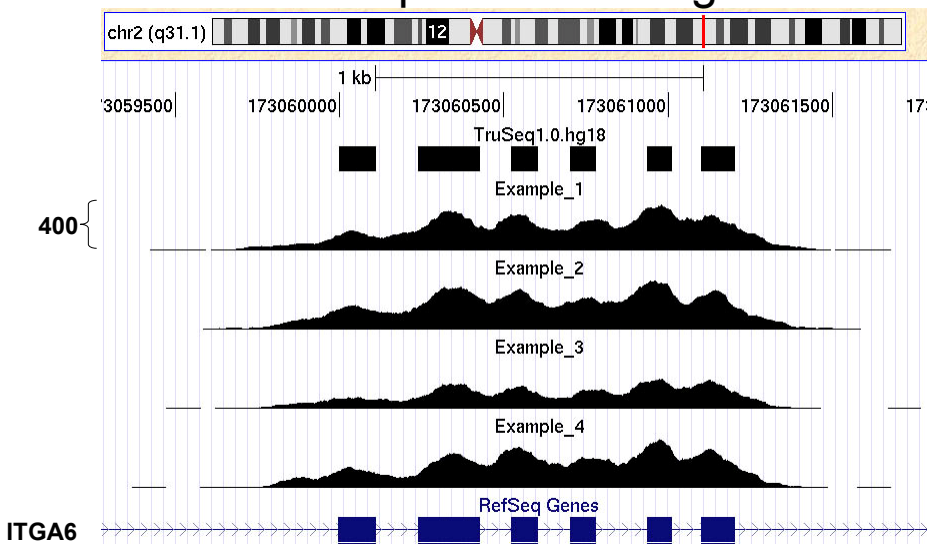


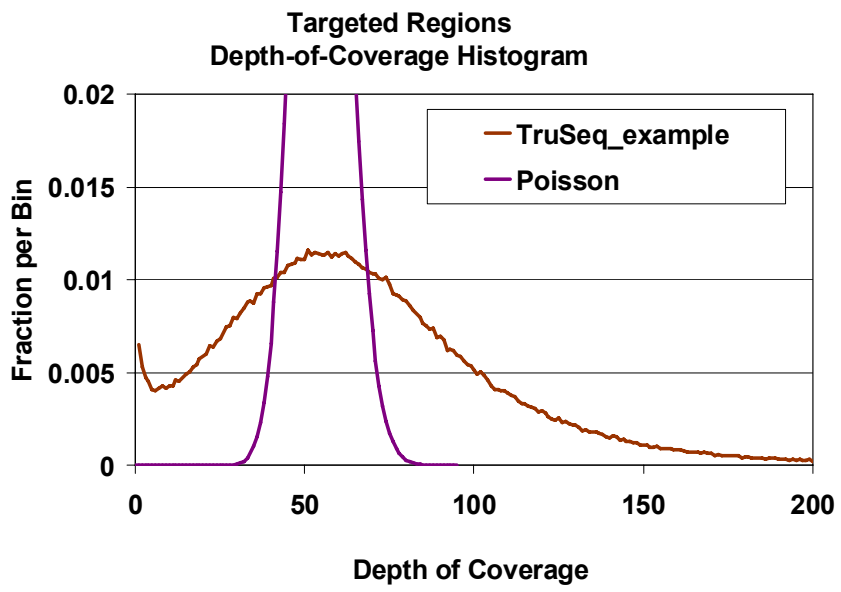
Total sequence per sample: ~10Gb from two lanes
Over 100X coverage of targeted regions

Read Depth-of-Coverage



Read Depth-of-Coverage

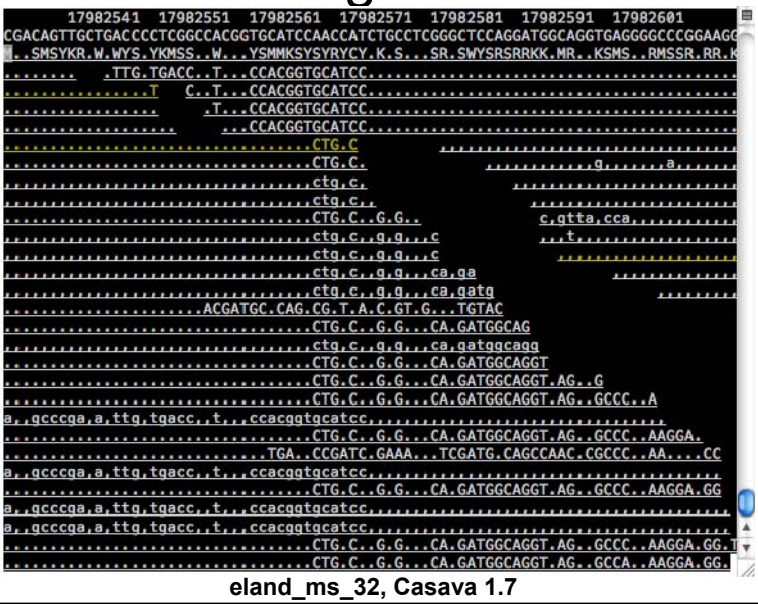




Refining the Alignment (diagCM)

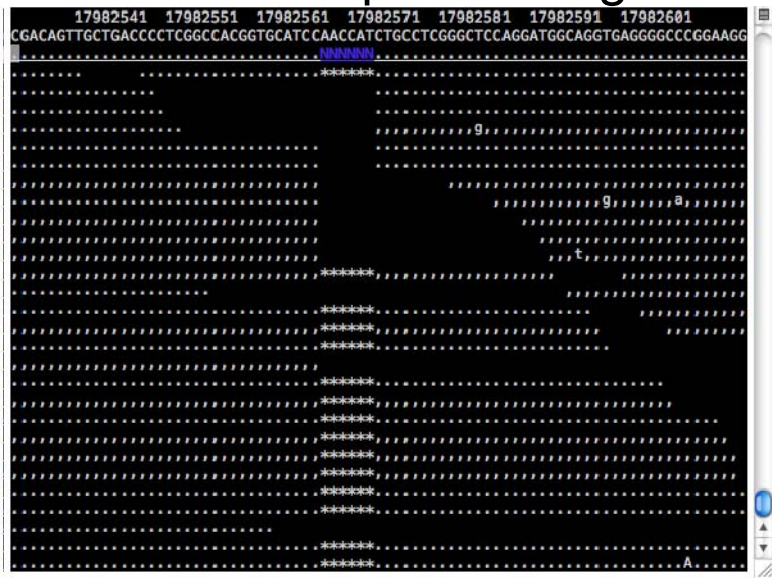
- ELAND is part of the standard pipeline
- ELAND accurately places reads in the correct genomic location
- Use cross-match, a Smith-Waterman aligner, to improve local alignment

ELAND Aligned Reads

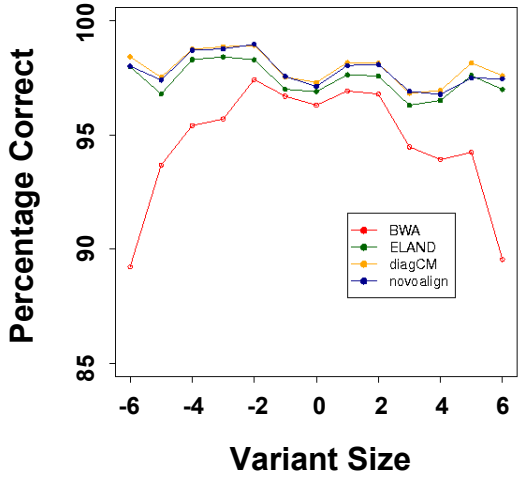


eland_ms_32, Casava 1.7

Cross-match Improved Alignment



Comparison of Aligners (simulated exome data)



Bayesian Genotype Calling: Most Probable Genotype

```
ACCTTCCTCGAGGCTAGCCTAGCCGGGGTGGGCAAGGCCTTCCGGGGGGGTAACCTCCTCCAAAACCTCCCTATGCCCCCCCAATTTTTTATATATATCCTCCTCCAA
     ..............................................A...
       ..............................................
         ............................................A.......
           ..........................................A.........
             ........................................A...........
                 ..................................................
                   ................................A.................
                     ..............................................
                       ............................A.......................
                           ......................A.........................
                             ....................A.........................
                               ..................A.............................
                                 ................A.................................
                                   ..............A.................................
                                     ............A.................................
                                       ..........A.................................
                                         ........A.................................
```

P(Genotype|Data)

AA	10^{-6}
AC	10^{-15}
AG	10^{-15}
AT	0.99999981
CC	10^{-75}
CG	10^{-75}
CT	10^{-60}
GG	10^{-75}
GT	10^{-60}
TT	10^{-54}

Convert to natural Log Probabilities and Sort

AT	0
AA	-14
AC	-34
AG	-34
TT	-124
CT	-138
GT	-138
CC	-172
CG	-172
GG	-172

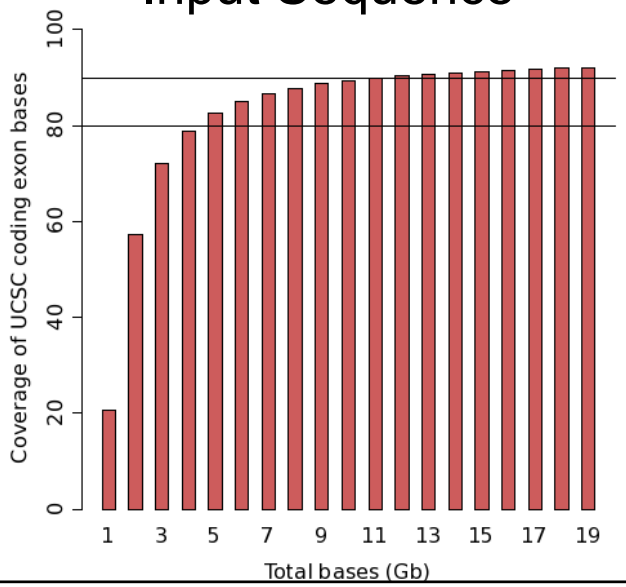
Difference between most probable and next most probable genotype

MPG Score = 14

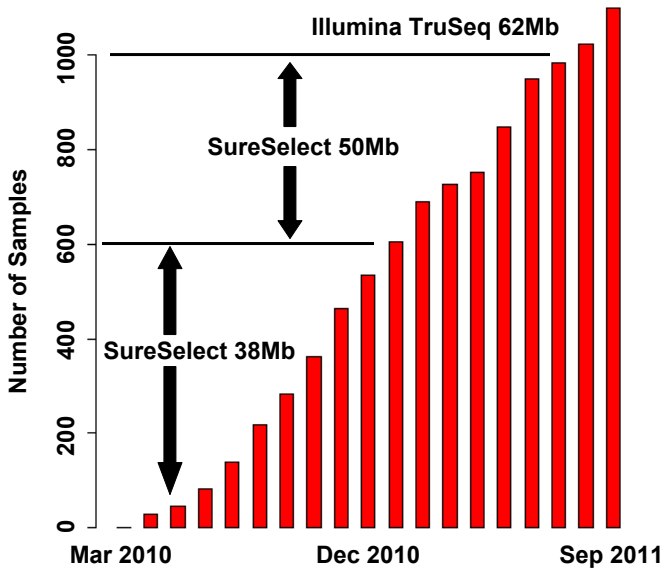
MPG of Haploid Regions

- Human autosomes are normally diploid
- MPG is designed to call two alleles for the autosomes, and X chromosome if the sample is from a female
- For samples from males, MPG is run in haploid mode on the non-sudoautosomal regions of X and Y
- Thus only testing for the four nucleotides

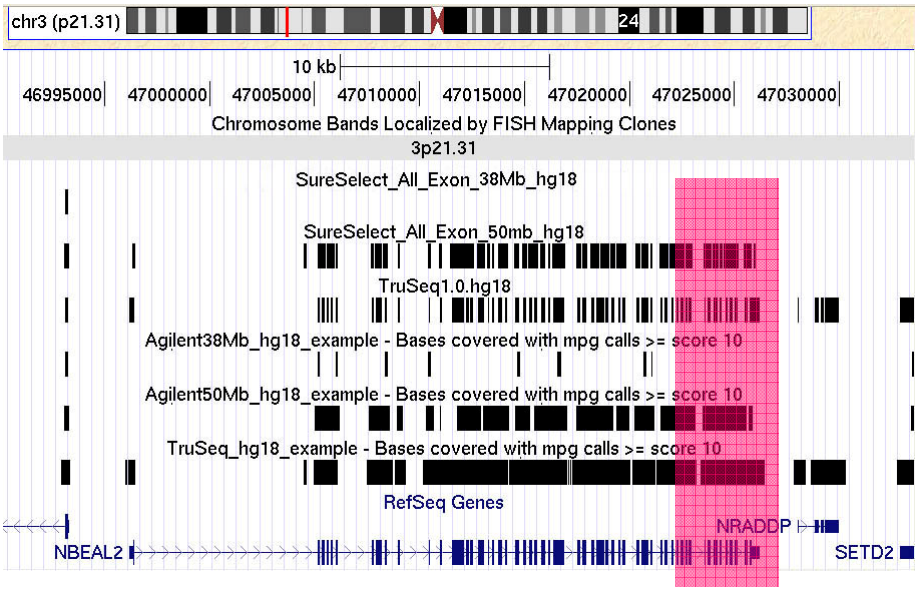
Exome Coverage versus Input Sequence



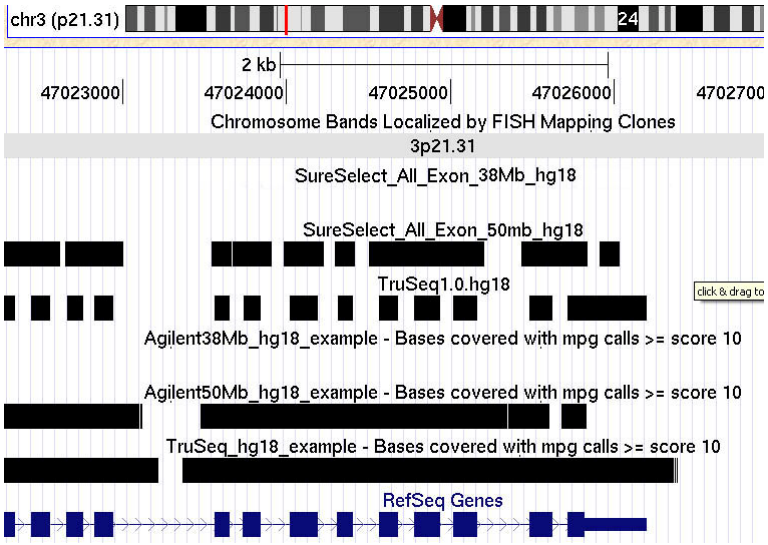
Whole Exomes Processed at NISC



NBEAL2



NBEAL2 5'



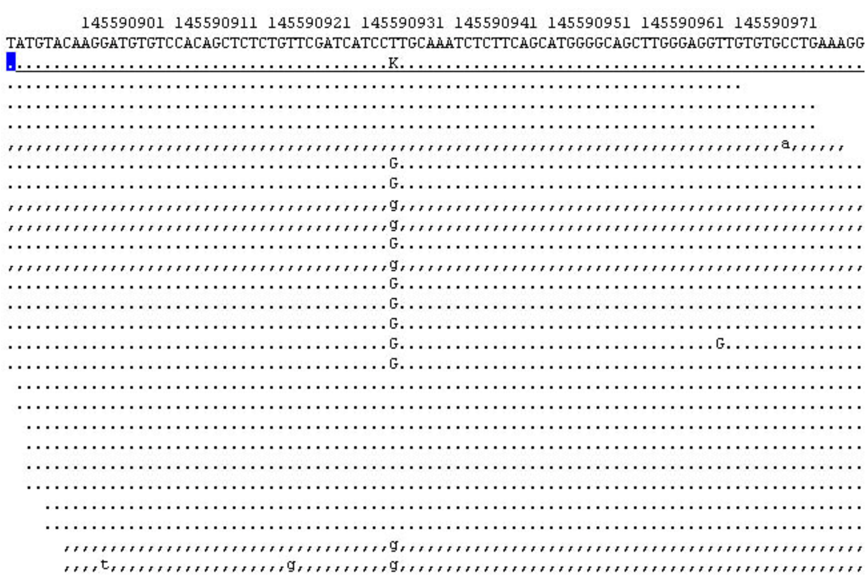
Exome Variation Statistics TruSeq 62Mb, Male Sample

Type	Total Genotype Calls	SNVs	Within-sample Heterozygosity
Total	133,047,403	142,361	0.00072
Auto	125,491,045	139,295	0.00076
chrX	6,842,299	2,600	NA
chrY	710,243	1,435	NA

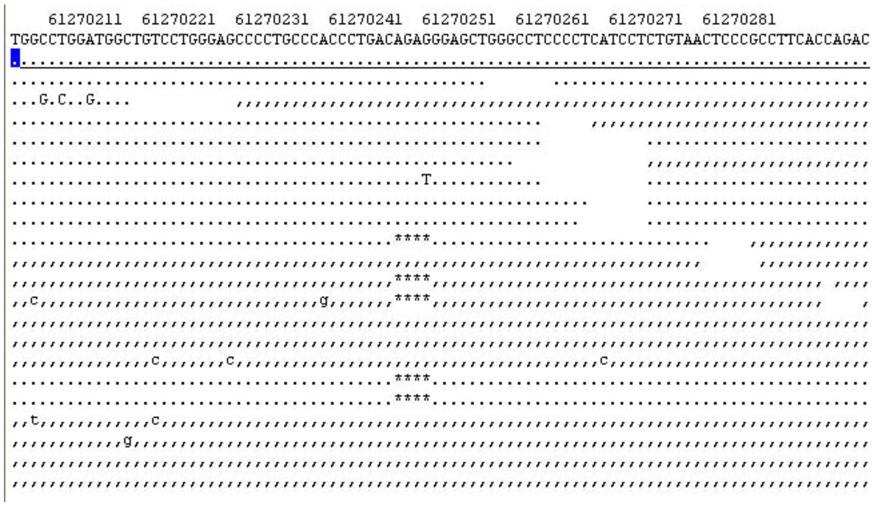
Exome Variation Statistics
TruSeq 62Mb, Female Sample

Type	Total Genotype Calls	SNVs	Within-sample Heterozygosity
Total	125,681,915	136,993	0.00075
Auto	120,559,746	132,616	0.00076
chrX	5,096,376	2,701	0.00034

Example Heterozygous SNV



Example Heterozygous Deletion



Coverage of MPG >= 10 Genotype Calls

	Total Raw Sequence	Aligned Sequence	Genotype calls CCDS	Genotype calls UCSC coding
SureSelect 38Mb	6.7 Gb	5.0 Gb (131x)	89%	74%
SureSelect 50Mb	10.5 Gb	6.1 Gb (122x)	89%	85%
TruSeq 62Mb	9.0 Gb	7.1 Gb (114x)	91%	89%
Whole Genome Shotgun	192 Gb	133 Gb (44x)	86%	83%

Genotype Concordance

	Total Agreement with Genotype Chip (CCDS)
Whole Genome Shotgun	99.908%
SureSelect 38Mb	99.910%
SureSelect 50Mb	99.857%
TruSeq 62Mb	99.865%

Whole Exome Sequencing

- Being applied to
  - Undiagnosed Diseases Program (100's of samples)
  - ClinSeq (>1000 samples)
  - Variety of other PI driven projects (e.g. cancer)
- Data generation rate per year
  - 200 exomes per GAiiX
  - 1200 exomes per HiSeq 2000
- Analysis results
  - Genotype data for 90% of consensus coding exon bases (CCDS)
  - Accuracy of genotype calls over 99.5%

Exome Sequencing Pipeline

Sample DNA Fragmentation

Illumina Library Preparation

Exome Enrichment

Cluster Generation

Sequencing and Basecalling

Sequence Read Alignment

Variation Detection

Variant Annotation and Working With Whole-Exome Data

- One sample produces > 100k variants
- One hundred samples gives rise to 600k or more
- How does one work with such large datasets?
- The next speaker, Dr. Jamie Teer, will address these next steps

Acknowledgements

NIH Intramural Sequencing Center

- Sequencing Operations
  - Bob Blakesley
  - Alice Young
  - Lab Staff
- Bioinformatics
  - Gerry Bouffard
  - Baishali Maskeri
  - Jenny McDowell
  - Meg Vemulapalli
- IT Linux Support
  - Jesse Becker
  - Matt Lesko
- Mullikin Lab
  - Nancy Hansen
  - Pedro Cruz
  - Praveen Cherukuri
- Biesecker Lab
  - Jamie Teer

http://research.nhgri.nih.gov/