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

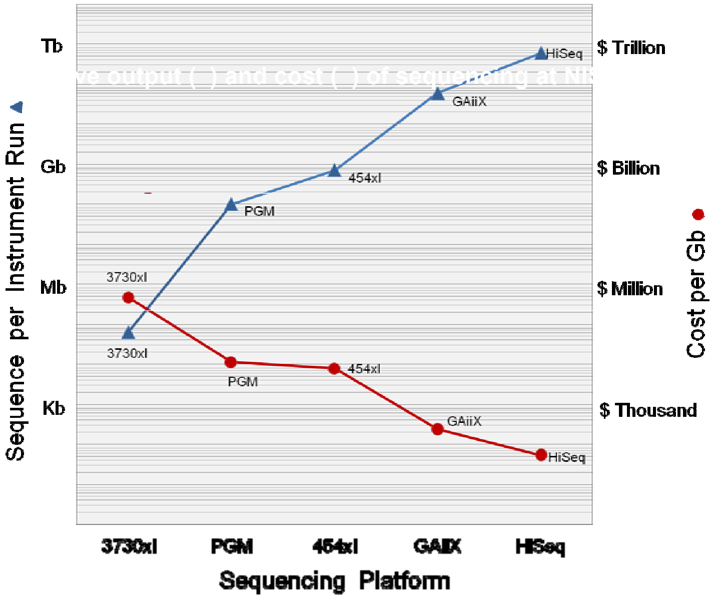
5625 Fishers Lane, Rockville MD

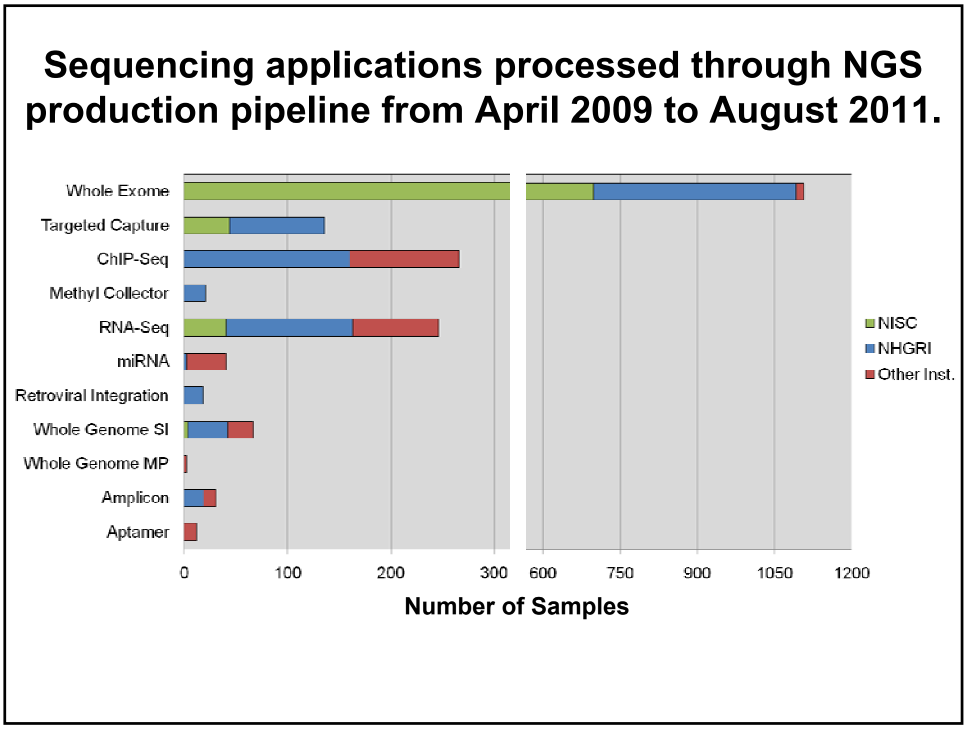
NISC Sequence Production

Feb. 2010

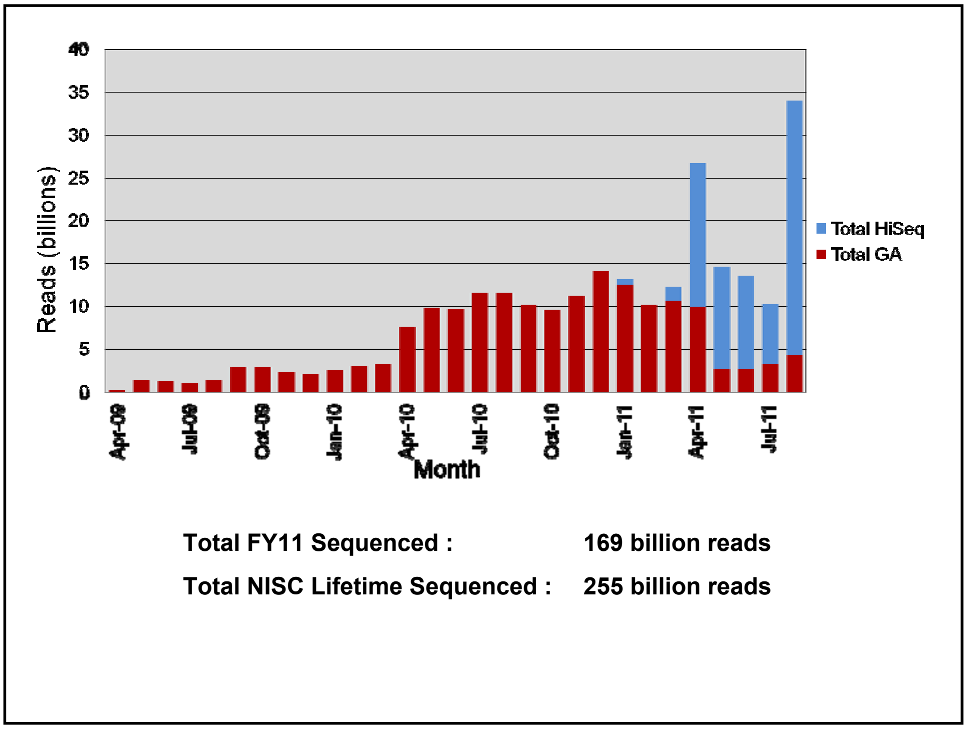
NISC Sequence Production

March 2011





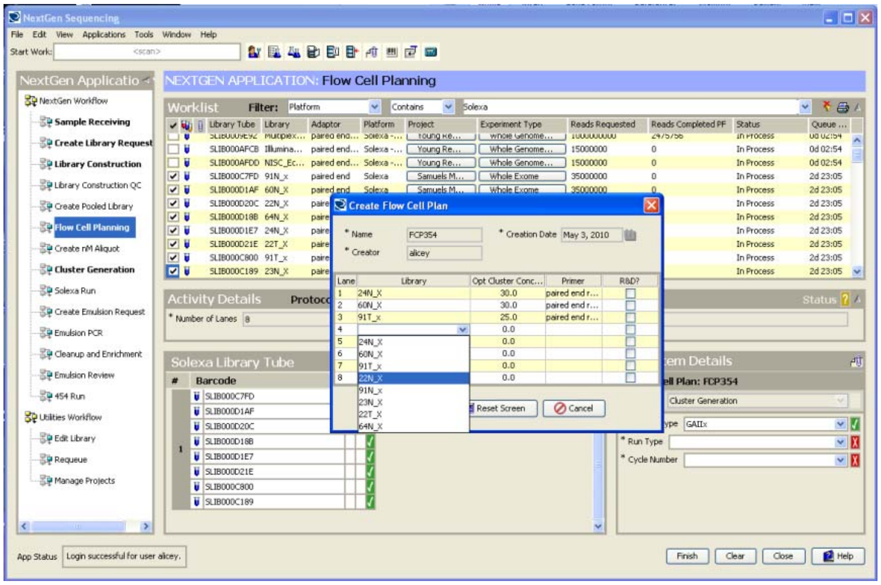
Sequencing applications processed through NGS production pipeline from April 2009 to August 2011.



Total FY11 Sequenced : **169 billion reads**

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

Cimarron Software based NextGen LIMS



NEXTGEN APPLICATION: Flow Cell Planning

Worklist Filter: Platform Contains Solexa

✓	Library Tube	Library	Adaptor	Platform	Project	Experiment Type	Reads Requested	Reads Completed PF	Status	Queue …
□	SLIB0009E92	Multiplex…	paired end…	Solexa -…	Young Re…	Whole Genome…	1000000000	2475756	In Process	0d 02:54
□	SLIB000AFCB	Illumina…	paired end…	Solexa -…	Young Re…	Whole Genome…	15000000	0	In Process	0d 02:54
□	SLIB000AFDD	NISC_Ec…	paired end…	Solexa -…	Young Re…	Whole Genome…	15000000	0	In Process	0d 02:54
☑	SLIB000C7FD	91N_x	paired end	Solexa	Samuels M…	Whole Exome	35000000	0	In Process	2d 23:05
☑	SLIB000D1AF	60N_X	paired end	Solexa	Samuels M…	Whole Exome	35000000	0	In Process	2d 23:05
☑	SLIB000D20C	22N_X	paire						In Process	2d 23:05
☑	SLIB000D18B	64N_X	paire						In Process	2d 23:05
☑	SLIB000D1E7	24N_X	paire						In Process	2d 23:05
☑	SLIB000D21E	22T_X	paire						In Process	2d 23:05
☑	SLIB000C800	91T_x	paire						In Process	2d 23:05
☑	SLIB000C189	23N_X	paire						In Process	2d 23:05

Create Flow Cell Plan

* Name FCP354 * Creation Date May 3, 2010

* Creator alicey

Lane	Library	Opt Cluster Conc…	Primer	R&D?
1	24N_X	30.0	paired end r…	□
2	60N_X	30.0	paired end r…	□
3	91T_x	25.0	paired end r…	□
4		0.0		□
5	24N_X	0.0		□
6	60N_X	0.0		□
7	91T_x	0.0		□
8	22N_X	0.0		□

Activity Details Protoco

* Number of Lanes 8

Status

Solexa Library Tube

#	Barcode
1	SLIB000C7FD
	SLIB000D1AF
	SLIB000D20C
	SLIB000D18B
	SLIB000D1E7
	SLIB000D21E
	SLIB000C800
	SLIB000C189

tem Details

ell Plan: FCP354

Cluster Generation

ype GAIIx

* Run Type

* Cycle Number

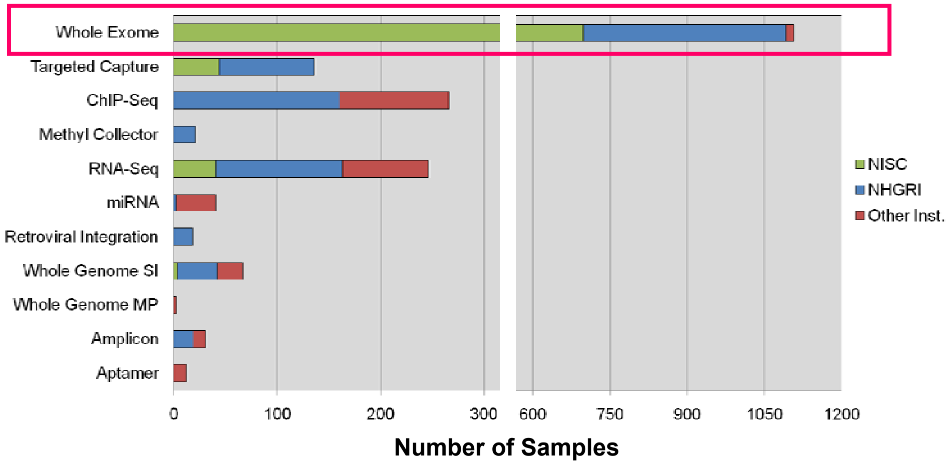
App Status Login successful for user alicey.

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

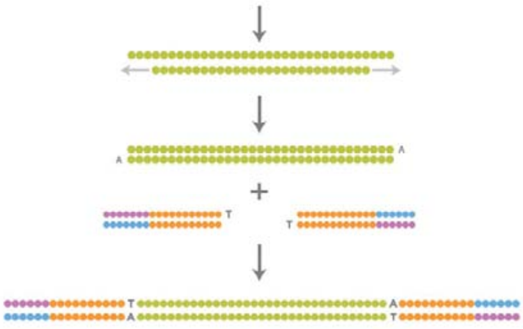
Cluster Generation

Sequencing and Basecalling

Sequence Read Alignment

Variation Detection

Library Preparation

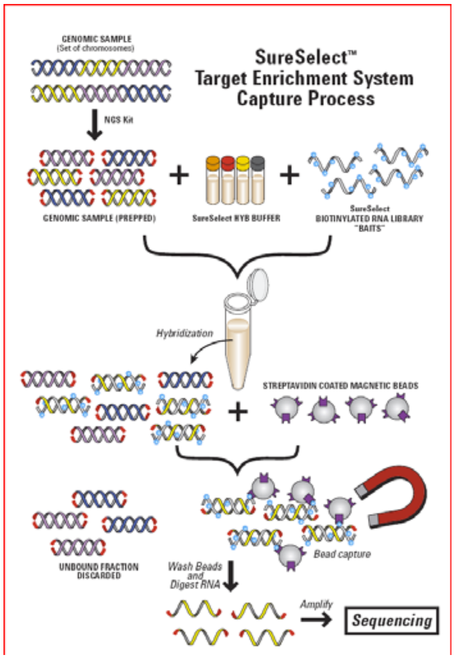


- A Fragment DNA
- B Repair ends Add A overhang
- C Ligate adapters
- D Select ligated DNA

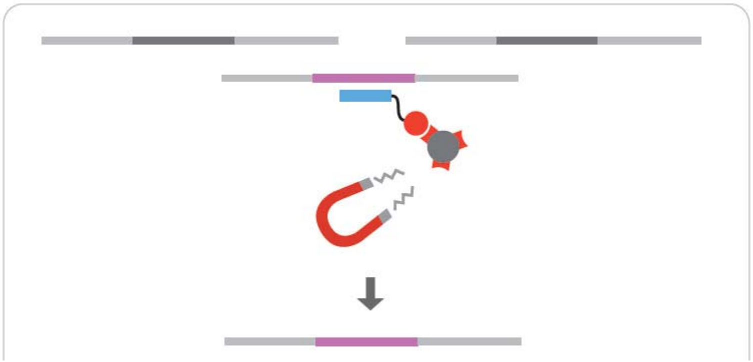
http://www.illumina.com/support/literature.ilmn

Agilent Technologies SureSelect method

Whole-exome kit 38Mb and 50Mb



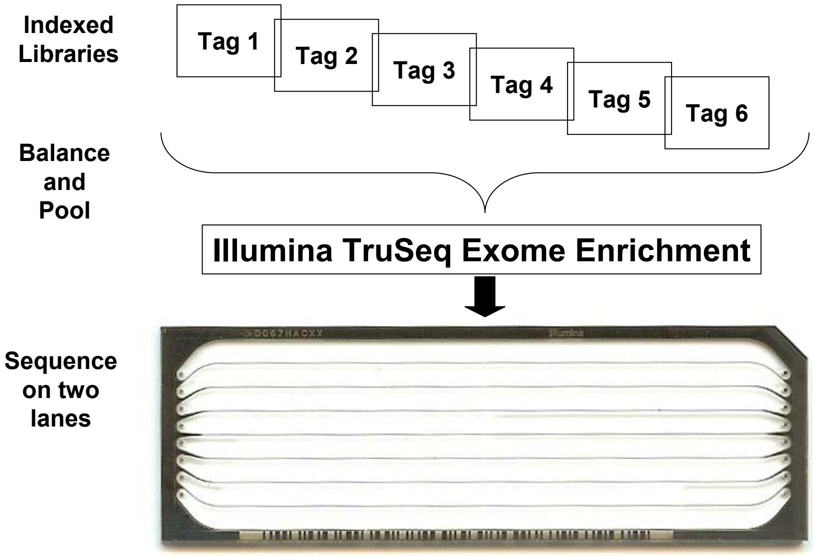
http://cp.literature.agilent.com/litweb/pdf/5990-3532EN.pdf



- Illumina TruSeq Exome Enrichment
- 62Mb of exome targeted

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

NISC Exome Process



Indexed Libraries

Tag 1

Tag 2

Tag 3

Tag 4

Tag 5

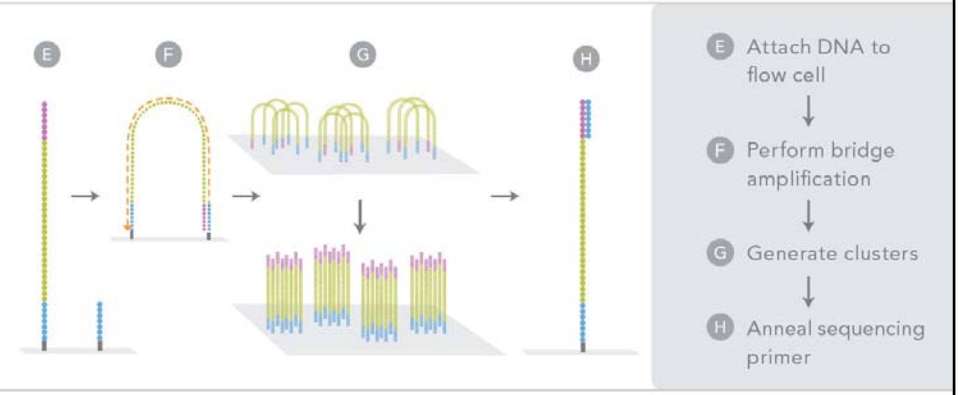
Tag 6

Balance and Pool

Illumina TruSeq Exome Enrichment

Sequence on two lanes

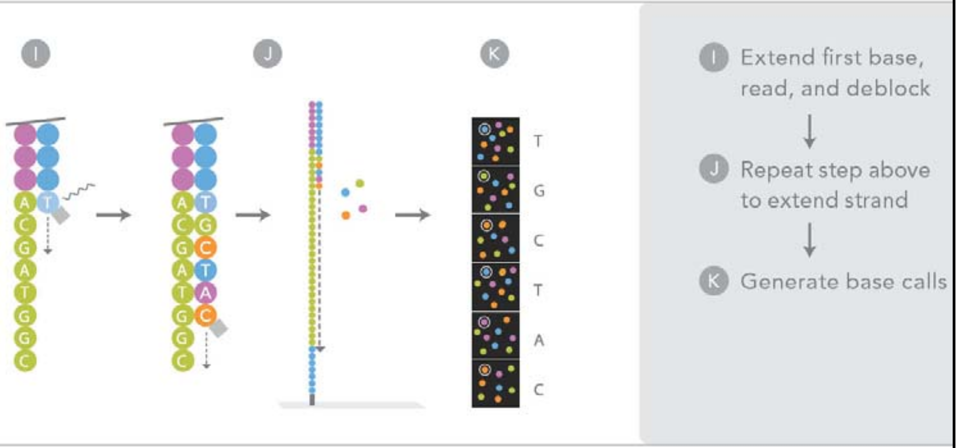
Cluster Generation



E Attach DNA to flow cell
↓
F Perform bridge amplification
↓
G Generate clusters
↓
H Anneal sequencing primer

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

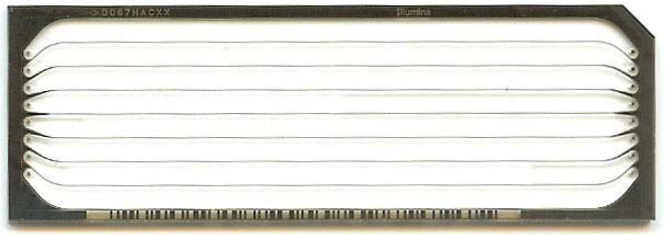
Sequencing



I Extend first base, read, and deblock
↓
J Repeat step above to extend strand
↓
K Generate base calls

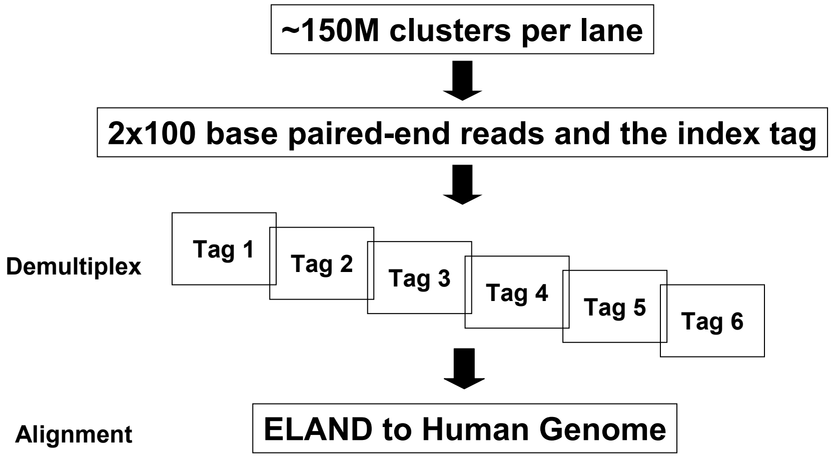
http://www.illumina.com/support/literature.ilmn

HiSeq Flow Cell



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

Sequencing and Data Processing



~150M clusters per lane

2x100 base paired-end reads and the index tag

Demultiplex

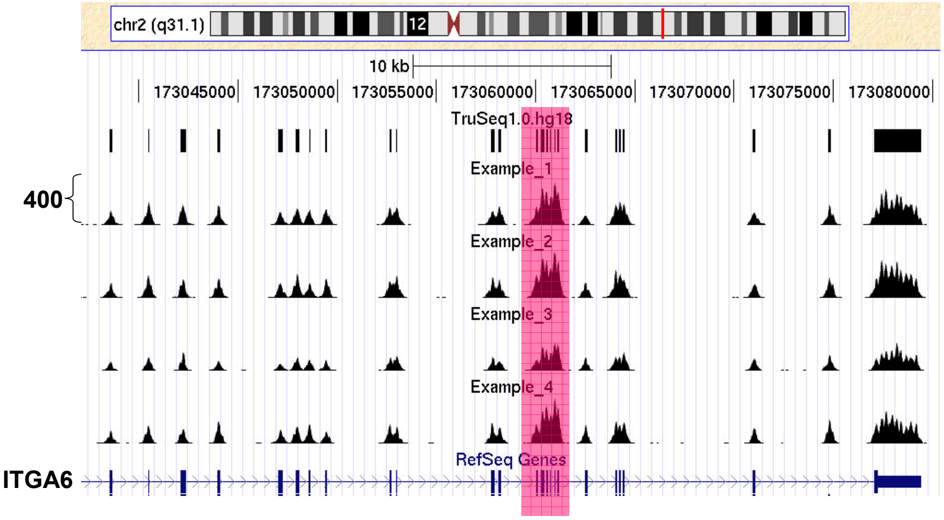
Tag 1 **Tag 2** **Tag 3** **Tag 4** **Tag 5** **Tag 6**

Alignment

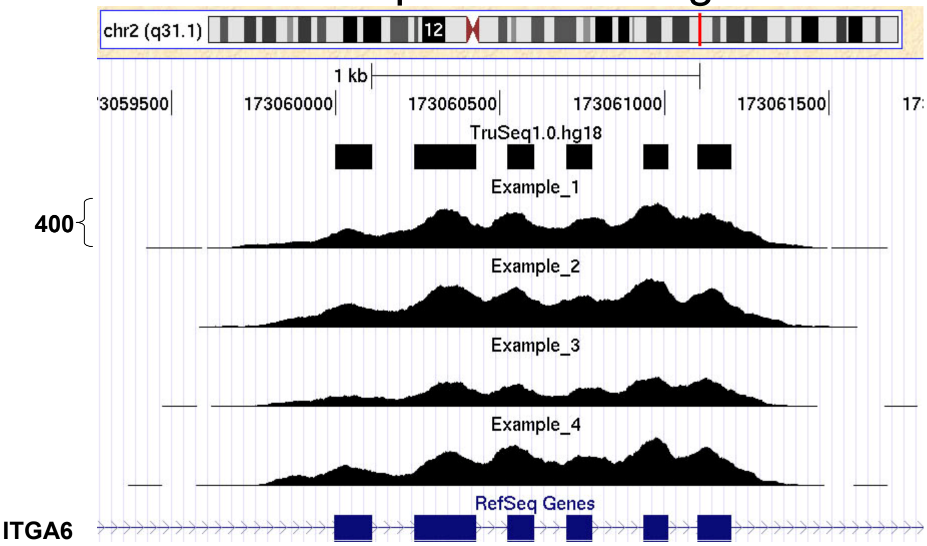
ELAND to Human Genome

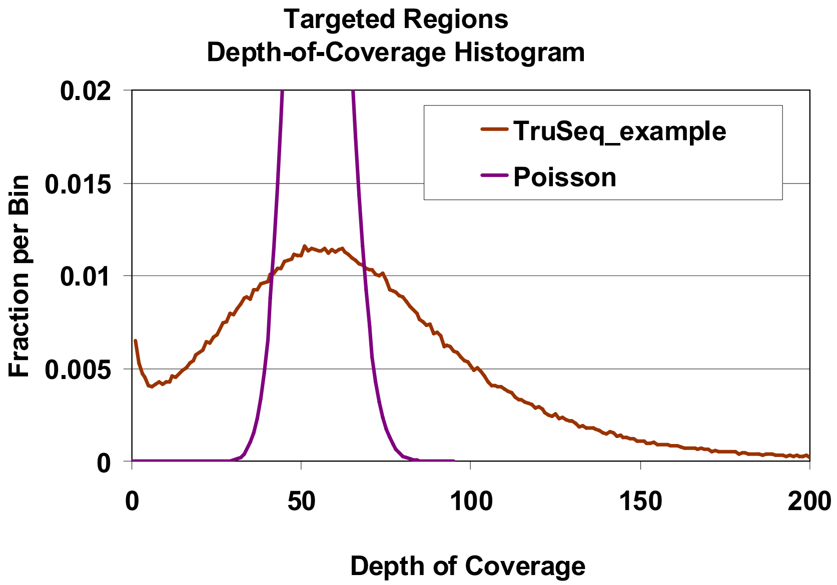
Total sequence per sample: ~10Gb from two lanes
Over 100X coverage of targeted regions

Read Depth-of-Coverage



Read Depth-of-Coverage

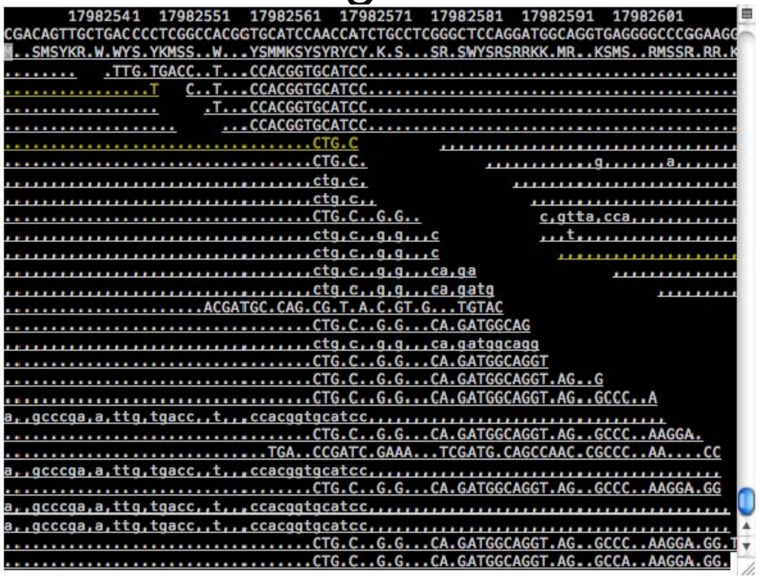




Refining the Alignment (diagCM)

- 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

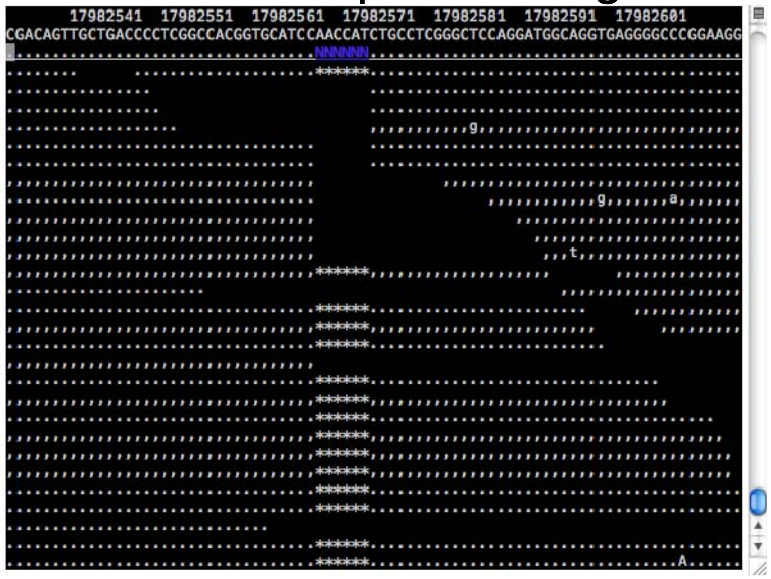
ELAND Aligned Reads



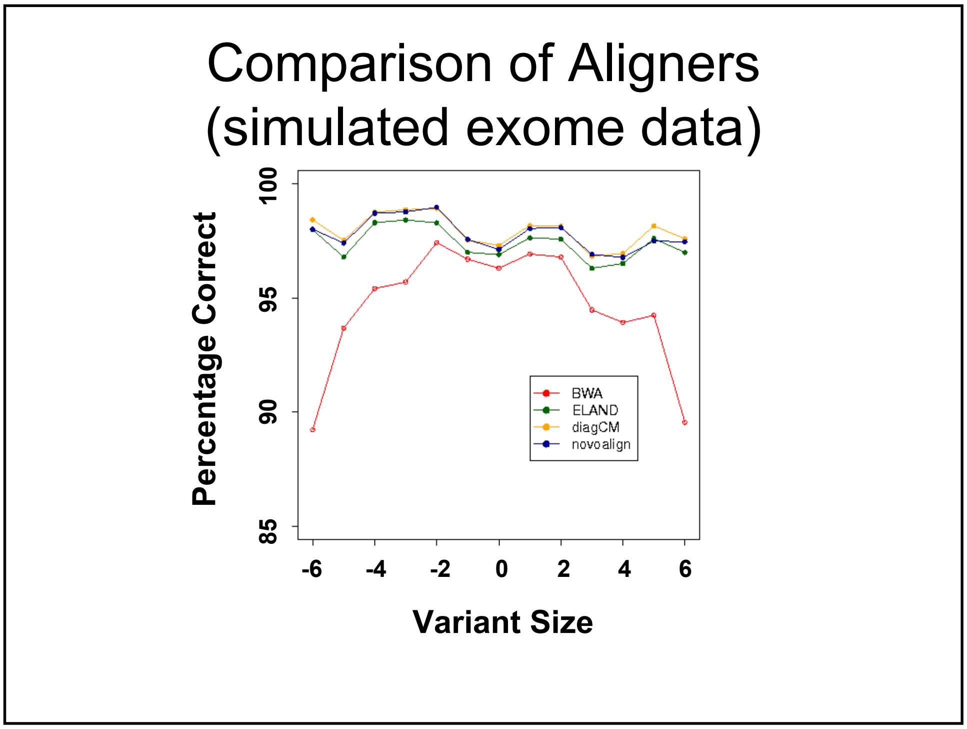
```
       17982541  17982551  17982561  17982571  17982581  17982591  17982601
CGACAGTTGCTGACCCCTCGGCCACGGTGCATCCAACCATCTGCCTCGGGCTCCAGGATGGCAGGTGAGGGGCCCGGAAGG
N..SMSYKR.W.WYS.YKMSS..W...YSMMKSYSYRYCY.K.S...SR.SWYSRSRRKK.MR..KSMS..RMSSR.RR.K
........   .TTG.TGACC..T...CCACGGTGCATCC.........................................
................T   C..T...CCACGGTGCATCC.........................................
.................     .T...CCACGGTGCATCC.........................................
...................     ...CCACGGTGCATCC.........................................
..................................CTG.C         ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................CTG.C.             ,,,,,,,,,,,,g,,,,,,,a,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,                ,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,                 ,,,,,,,,,,,,,,,,,,,,,,,
..................................CTG.C..G.G..             c,gtta,cca,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,g,g,,,c           ,,,t,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,g,g,,,c             ,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,g,g,,,ca,ga               ,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,g,g,,,ca,gatg                  ,,,,,,,,,
......................ACGATGC.CAG.CG.T.A.C.GT.G...TGTAC
..................................CTG.C..G.G...CA.GATGGCAG
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,ctg,c,,g,g,,,ca,gatggcagg
..................................CTG.C..G.G...CA.GATGGCAGGT
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..G
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..GCCC..A
a,,gcccga,a,ttg,tgacc,,t,,,ccacggtgcatcc,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..GCCC..AAGGA.
.............................TGA..CCGATC.GAAA...TCGATG.CAGCCAAC.CGCCC..AA....CC
a,,gcccga,a,ttg,tgacc,,t,,,ccacggtgcatcc,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..GCCC..AAGGA.GG
a,,gcccga,a,ttg,tgacc,,t,,,ccacggtgcatcc,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
a,,gcccga,a,ttg,tgacc,,t,,,ccacggtgcatcc,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..GCCC..AAGGA.GG.T
..................................CTG.C..G.G...CA.GATGGCAGGT.AG..GCCA..AAGGA.GG.
```

eland_ms_32, Casava 1.7

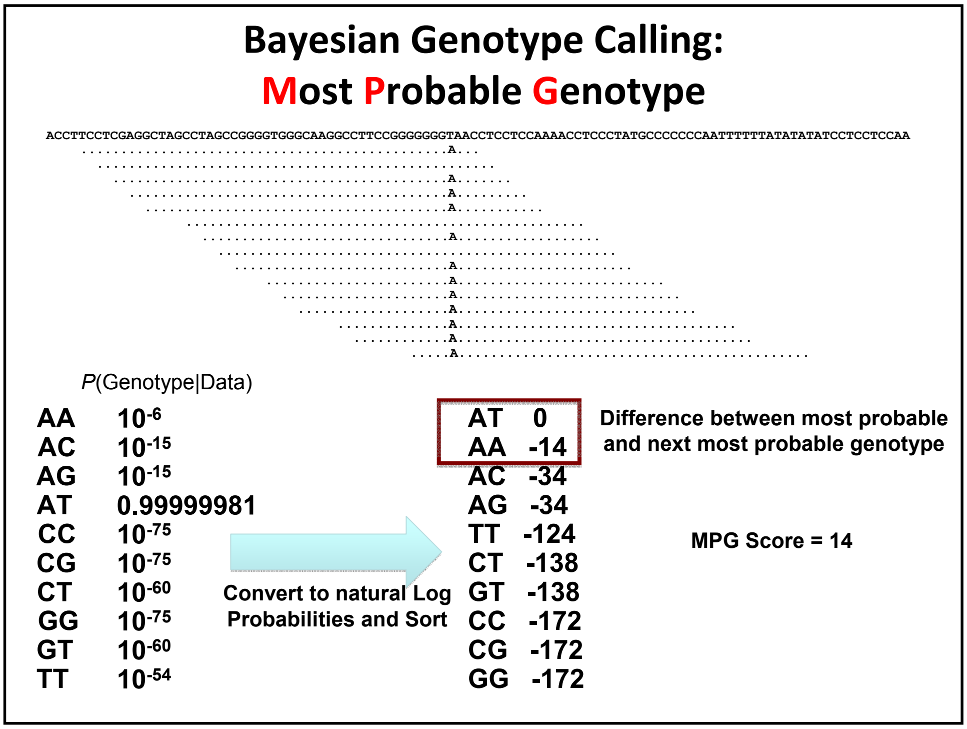
Cross-match Improved Alignment



```
       17982541  17982551  17982561  17982571  17982581  17982591  17982601
CGACAGTTGCTGACCCCTCGGCCACGGTGCATCCAACCATCTGCCTCGGGCTCCAGGATGGCAGGTGAGGGGCCCGGAAGG
..................................NNNNNN.........................................
........      ....................******.........................................
................                        .........................................
.................                       .........................................
...................                     ,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................      .........................................
..................................      .........................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,              ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................                   ,,,,,,,,,,,,g,,,,,,,a,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,                      ,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,                        ,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,                         ,,,t,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,       ,,,,,,,,,,,,,,
......................                                       ,,,,,,,,,,,,,,,,,,,,
..................................******........................     ,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,,,,,,       ,,,,,,,,,
..................................******..........................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................******................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................******......................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,******,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
..................................******.........................................
..................................******.........................................
.............................
..................................******.........................................
..................................******..................................A......
```



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.............................................
```

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

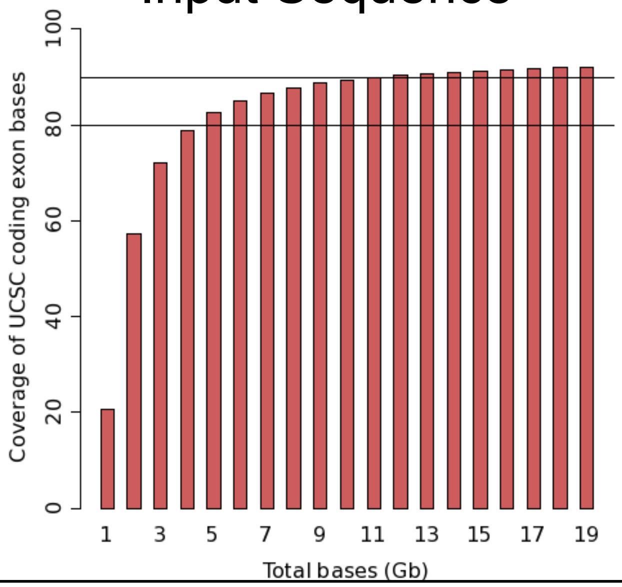
Difference between most probable and next most probable genotype

MPG Score = 14

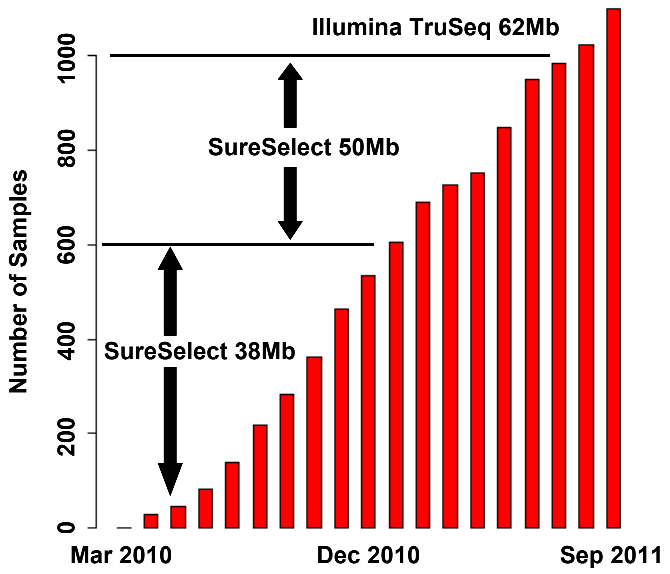
MPG of Haploid Regions

- 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

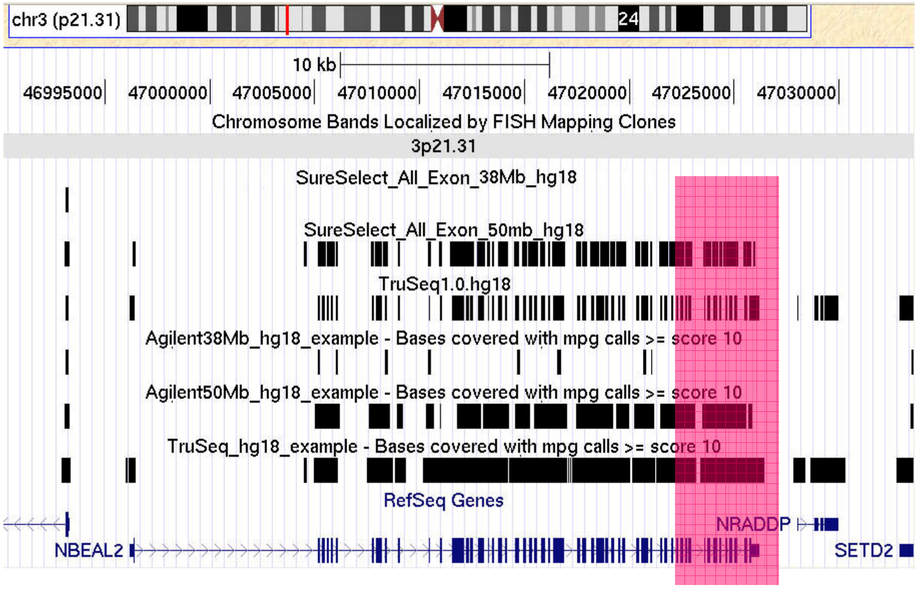
Exome Coverage versus Input Sequence



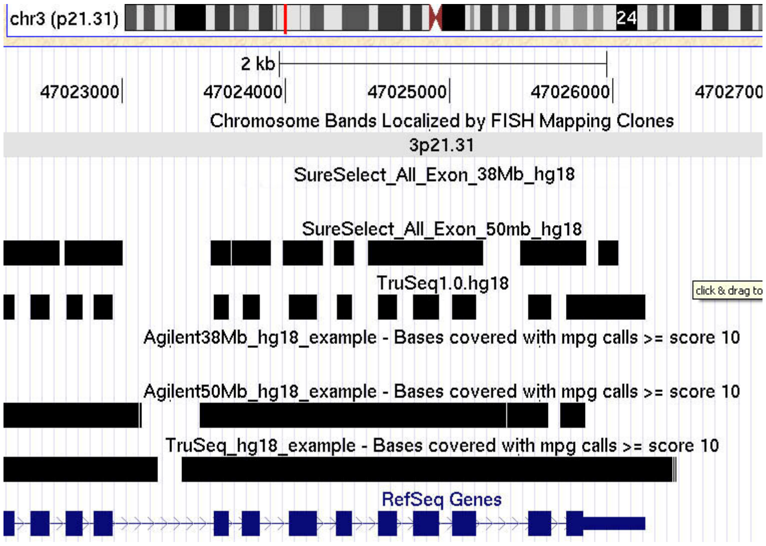
Whole Exomes Processed at NISC



NBEAL2


chr3 (p21.31)
24
10 kb
46995000
47000000
47005000
47010000
47015000
47020000
47025000
47030000
Chromosome Bands Localized by FISH Mapping Clones
3p21.31
SureSelect_All_Exon_38Mb_hg18
SureSelect_All_Exon_50mb_hg18
TruSeq1.0.hg18
Agilent38Mb_hg18_example - Bases covered with mpg calls >= score 10
Agilent50Mb_hg18_example - Bases covered with mpg calls >= score 10
TruSeq_hg18_example - Bases covered with mpg calls >= score 10
RefSeq Genes
NRADDP
NBEAL2
SETD2

NBEAL2 5’


chr3 (p21.31)
24
2 kb
47023000
47024000
47025000
47026000
4702700
Chromosome Bands Localized by FISH Mapping Clones
3p21.31
SureSelect_All_Exon_38Mb_hg18
SureSelect_All_Exon_50mb_hg18
TruSeq1.0.hg18
click & drag to
Agilent38Mb_hg18_example - Bases covered with mpg calls >= score 10
Agilent50Mb_hg18_example - Bases covered with mpg calls >= score 10
TruSeq_hg18_example - Bases covered with mpg calls >= score 10
RefSeq Genes

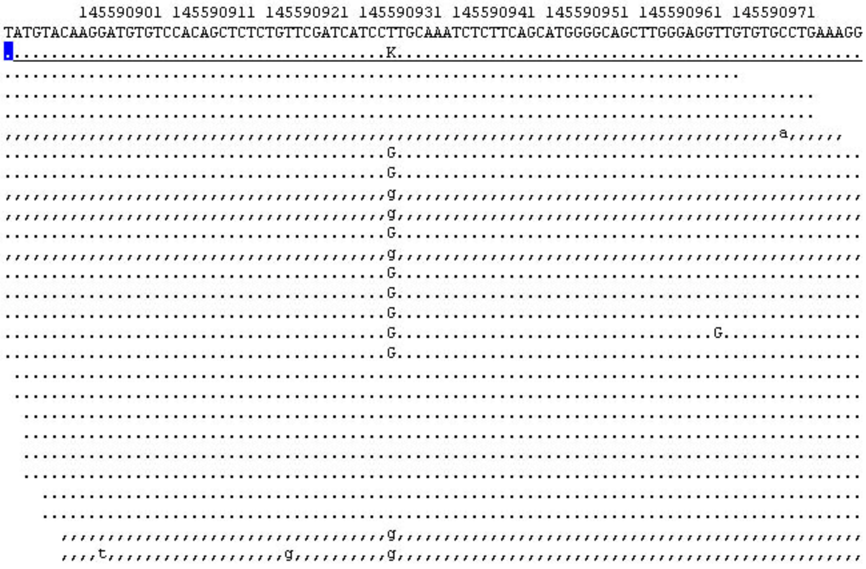
Exome Variation Statistics TruSeq 62Mb, Male Sample

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

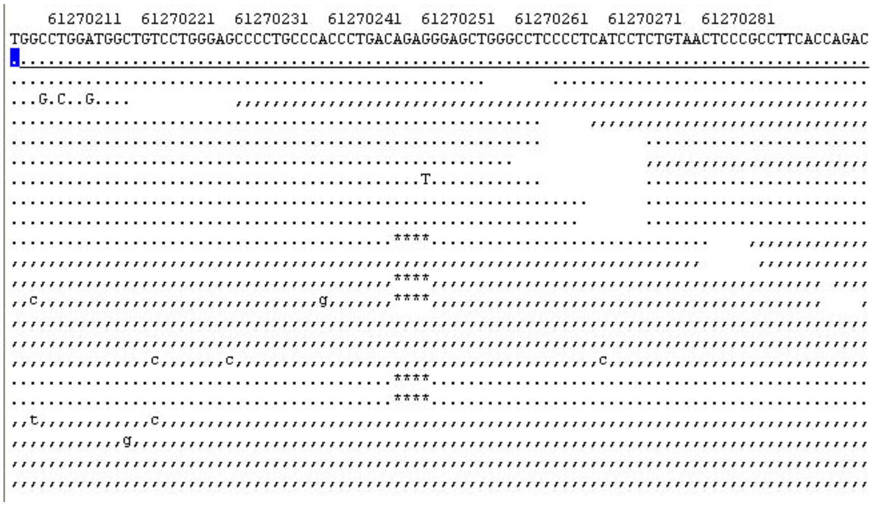
Type	Total Genotype Calls	SNVs	Within-sample Heterozygosity
Total	125,681,915	136,993	0.00075
Auto	120,559,746	132,616	0.00076
chrX	5,096,376	2,701	0.00034

Example Heterozygous SNV



```
        145590901 145590911 145590921 145590931 145590941 145590951 145590961 145590971
TATGTACAAGGATGTGTCCACAGCTCTCTGTTCGATCATCCTTGCAAATCTCTTCAGCATGGGGCAGCTTGGGAGGTTGTGTGCCTGAAAGG
.........................................K..................................................
...............................................................................
.......................................................................................
.......................................................................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,a,,,,,,
.........................................G..................................................
.........................................G..................................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................G..................................................
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................G..................................................
.........................................G..................................................
.........................................G..................................................
.........................................G..................................G...............
.........................................G..................................................
 ...........................................................................................
 ...........................................................................................
  ..........................................................................................
  ..........................................................................................
  ..........................................................................................
    ........................................................................................
    ........................................................................................
      ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
      ,,,,t,,,,,,,,,,,,,,,,,,,g,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

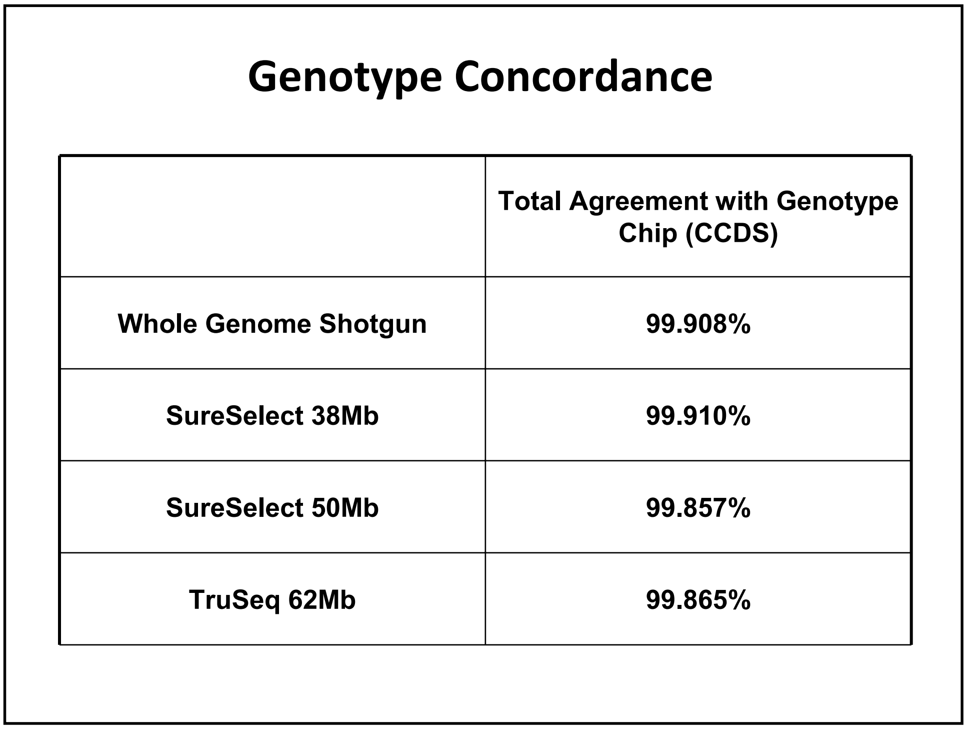
Example Heterozygous Deletion



```
    61270211  61270221  61270231  61270241  61270251  61270261  61270271  61270281
TGGCCTGGATGGCTGTCCTGGGAGCCCCTGCCCACCCTGACAGAGGGAGCTGGGCCTCCCCTCATCCTCTGTAACTCCCGCCTTCACCAGAC
............................................................................................
...................................................       ..................................
...G.C..G....           ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................................     ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................................           ........................
......................................................              ,,,,,,,,,,,,,,,,,,,,,,,,
............................................T............           ........................
..............................................................      ........................
.............................................................       ........................
.........................................****..............................    ,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,      ,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,****,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,
,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,g,,,,,,,****,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,    ,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,c,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.........................................****...............................................
.........................................****...............................................
,,t,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,g,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

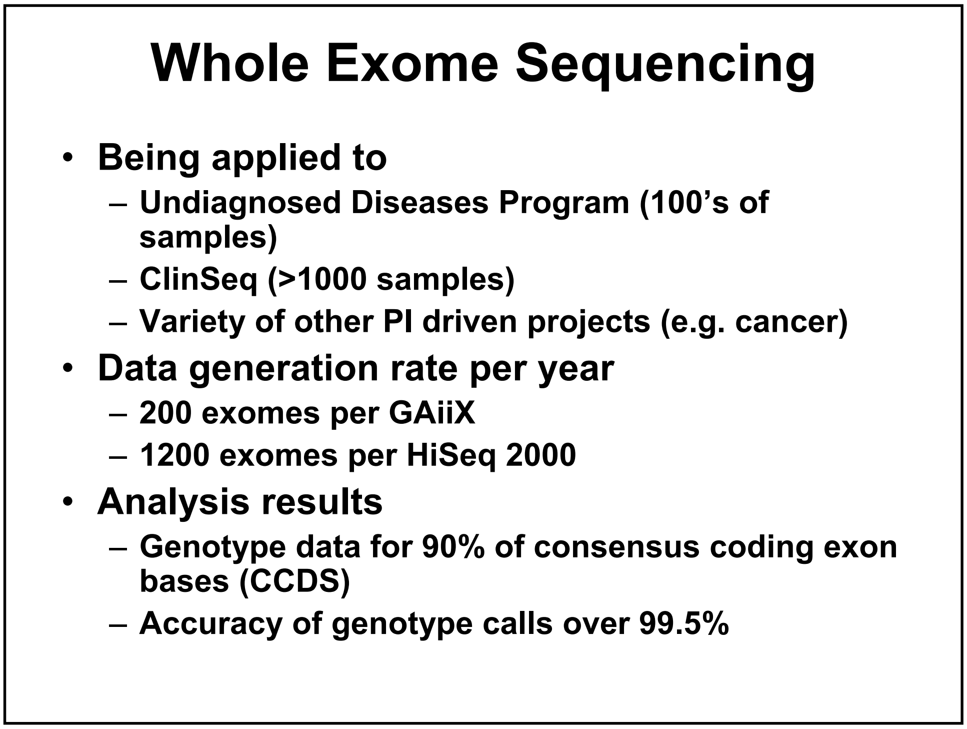
Coverage of MPG >= 10 Genotype Calls

	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/