NCI
April 2012

Comprehensive Detection of Variation in Thousands of Whole Genome Sequences

Stephen E. Lincoln
VP, Scientific Applications, Complete Genomics
slincoln@completegenomics.com

Early Examples of Complete Genomics Whole Genome Sequencing

Somatic Mutations in Cancer (Genentech)

- Compared NSCLC Tumor Resection to matched Normal
- ~50,000 Somatic SNPs at >90% validation rate
- 79 Somatic Structural Variations at a 66% validation rate
- **Finding: 1 Point Mutation per 3 Cigarettes smoked**

Lee et al., Nature 2010

Family of Four with Multiple Inherited Diseases (ISB)

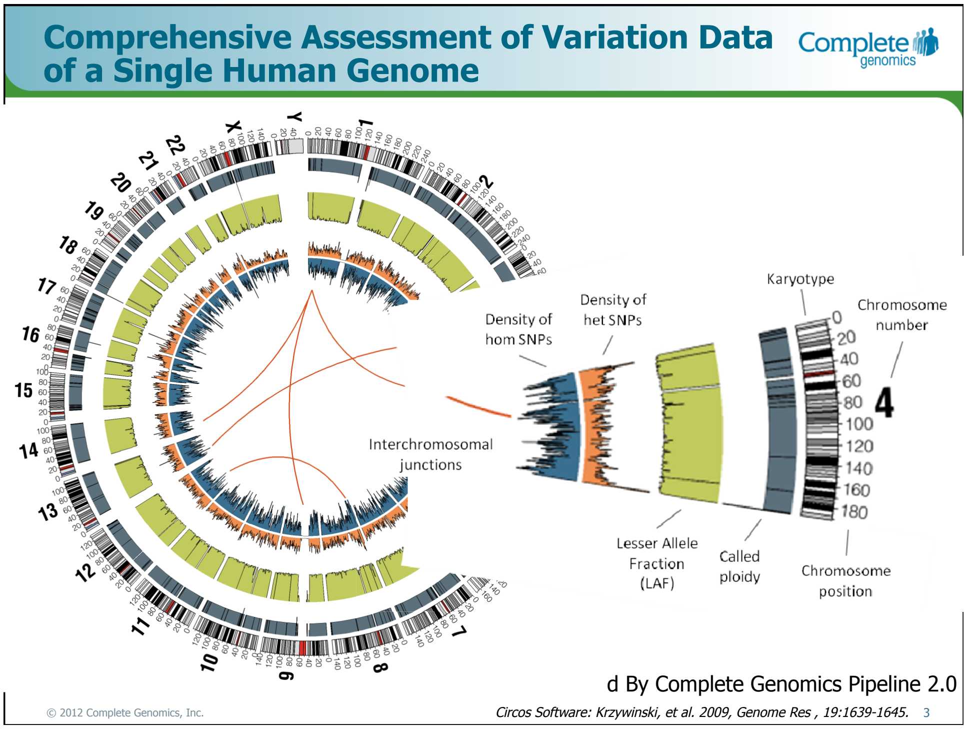
- **Found Both Causal Loci,** independently confirmed on an independent sequencing platform
- Measured ***de novo*** **Mutation Rate** in Meioses: 1.1 x 10^{-8}
- Benchmarked accuracy of the Complete platform

Roach et al., Science 2010

Affected Individual with Idiopathic Disease (UTSW)

- 11-Month Old with Severe Hypercholesterolemia
- Blood Test and Traditional DNA tests failed to identify cause
- **Genome sequencing showed required protein absent which had been missed by other genetic and biochemical tests**

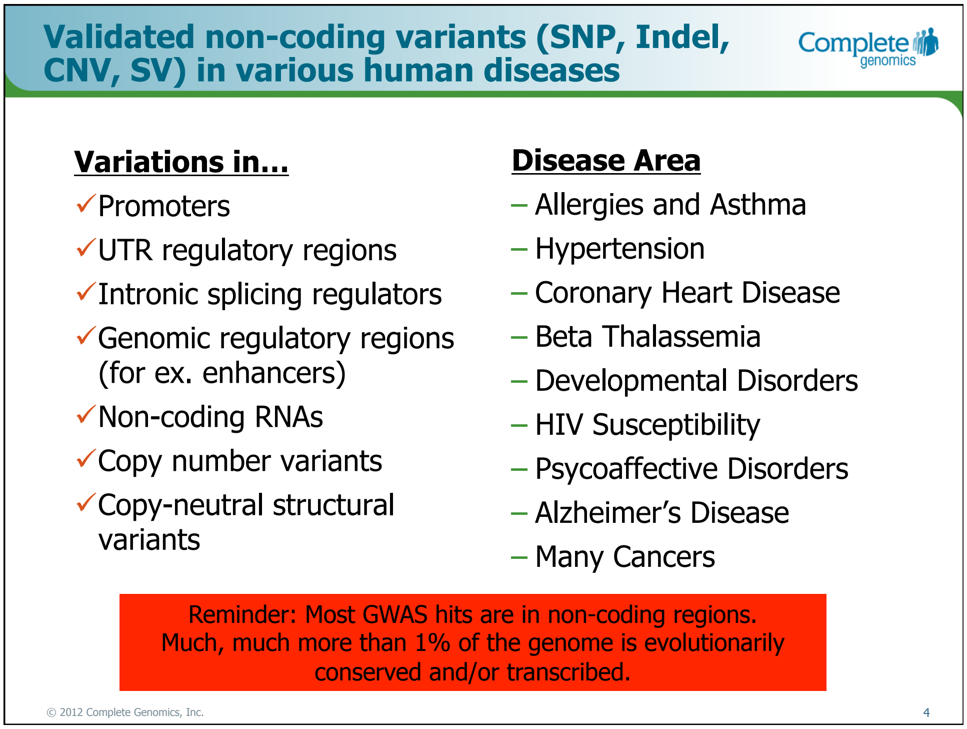
Rios et al., HMG 2010



Comprehensive Assessment of Variation Data of a Single Human Genome

d By Complete Genomics Pipeline 2.0

Circos Software: Krzywinski, et al. 2009, Genome Res , 19:1639-1645.



Validated non-coding variants (SNP, Indel, CNV, SV) in various human diseases

Complete genomics

Variations in…

✓Promoters

✓UTR regulatory regions

✓Intronic splicing regulators

✓Genomic regulatory regions (for ex. enhancers)

✓Non-coding RNAs

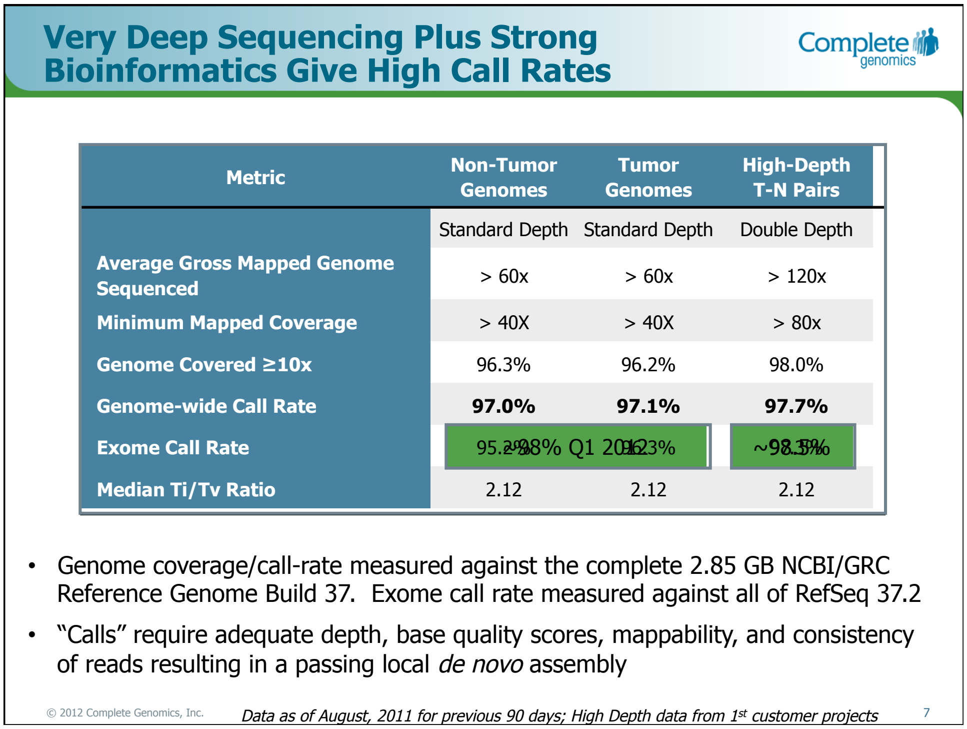
✓Copy number variants

✓Copy-neutral structural variants

Disease Area

– Allergies and Asthma
– Hypertension
– Coronary Heart Disease
– Beta Thalassemia
– Developmental Disorders
– HIV Susceptibility
– Psycoaffective Disorders
– Alzheimer’s Disease
– Many Cancers

Reminder: Most GWAS hits are in non-coding regions. Much, much more than 1% of the genome is evolutionarily conserved and/or transcribed.



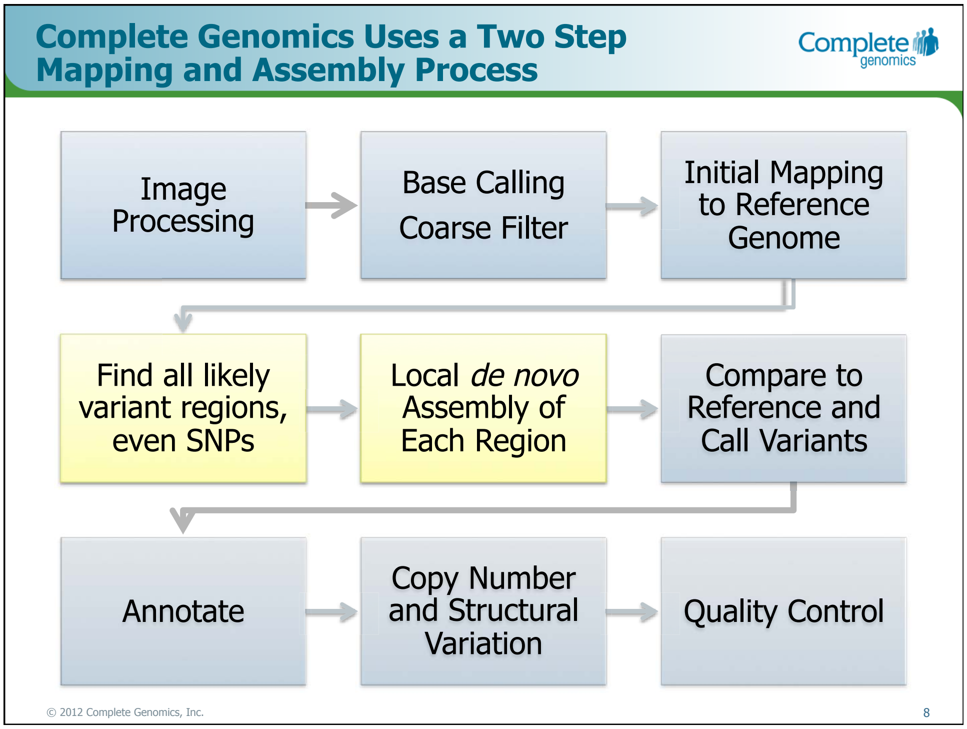
Very Deep Sequencing Plus Strong Bioinformatics Give High Call Rates

Complete genomics

Metric	Non-Tumor Genomes	Tumor Genomes	High-Depth T-N Pairs
	Standard Depth	Standard Depth	Double Depth
Average Gross Mapped Genome Sequenced	> 60x	> 60x	> 120x
Minimum Mapped Coverage	> 40X	> 40X	> 80x
Genome Covered ≥10x	96.3%	96.2%	98.0%
Genome-wide Call Rate	**97.0%**	**97.1%**	**97.7%**
Exome Call Rate	95.2% ~98% Q1 2012	96.3%	~98.5%
Median Ti/Tv Ratio	2.12	2.12	2.12

- Genome coverage/call-rate measured against the complete 2.85 GB NCBI/GRC Reference Genome Build 37. Exome call rate measured against all of RefSeq 37.2
- "Calls" require adequate depth, base quality scores, mappability, and consistency of reads resulting in a passing local *de novo* assembly

Data as of August, 2011 for previous 90 days; High Depth data from 1st customer projects



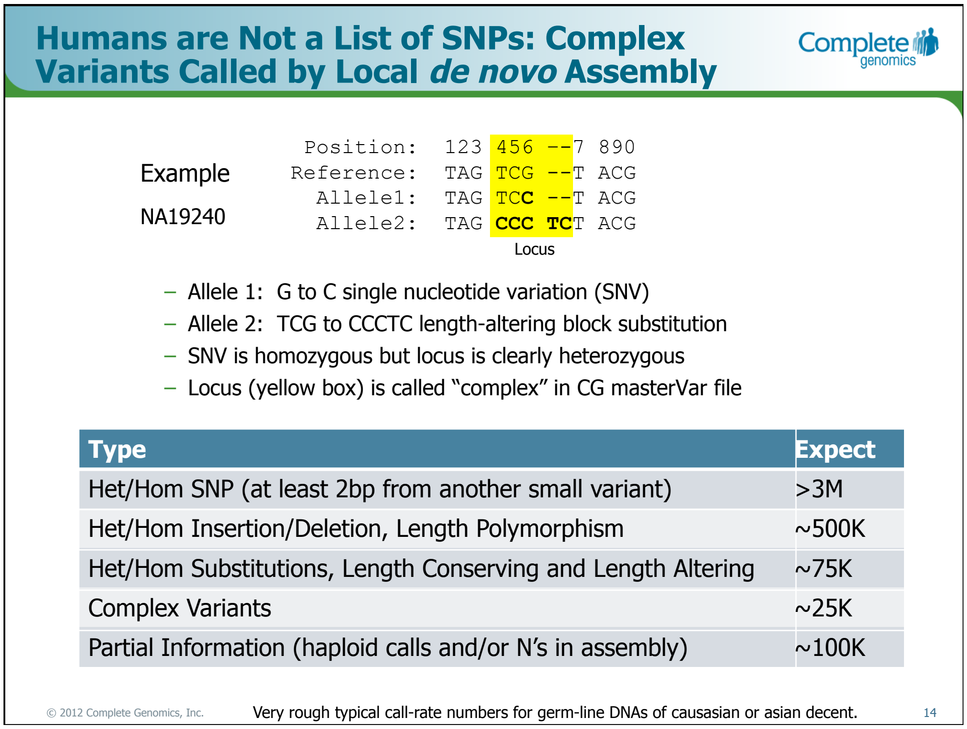
Complete Genomics Uses a Two Step Mapping and Assembly Process

Complete genomics

Image Processing → Base Calling Coarse Filter → Initial Mapping to Reference Genome

→ Find all likely variant regions, even SNPs → Local *de novo* Assembly of Each Region → Compare to Reference and Call Variants

→ Annotate → Copy Number and Structural Variation → Quality Control



Humans are Not a List of SNPs: Complex Variants Called by Local *de novo* Assembly

Complete genomics

Example

NA19240

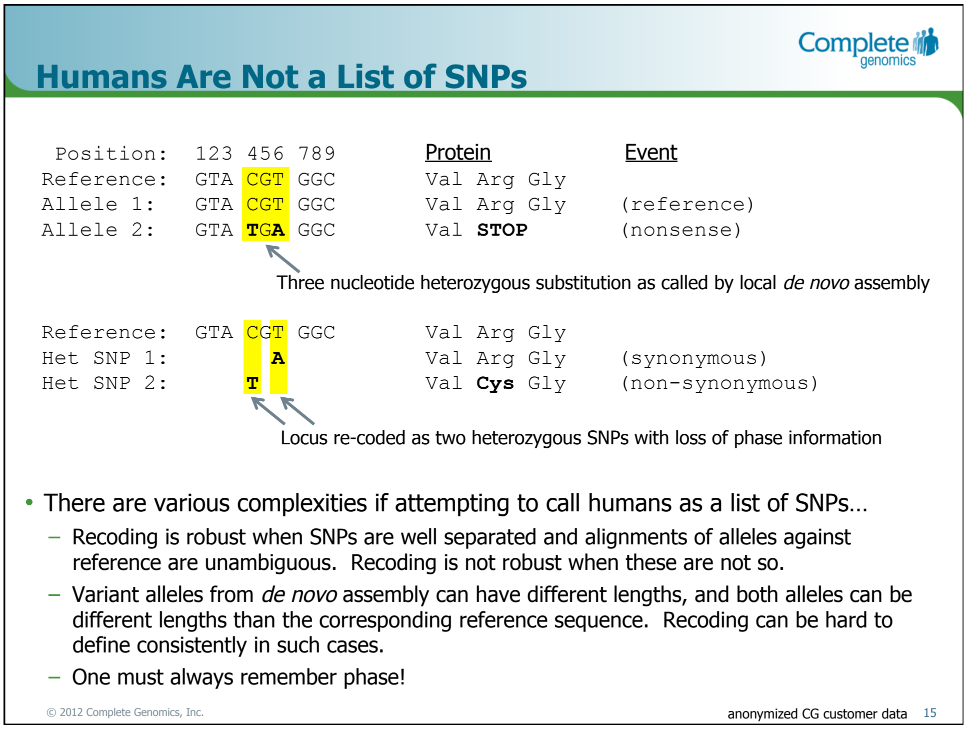
```
 Position:  123 456 --7 890
Reference:  TAG TCG --T ACG
  Allele1:  TAG TCC --T ACG
  Allele2:  TAG CCC TCT ACG
                 Locus
```

- Allele 1: G to C single nucleotide variation (SNV)
- Allele 2: TCG to CCCTC length-altering block substitution
- SNV is homozygous but locus is clearly heterozygous
- Locus (yellow box) is called "complex" in CG masterVar file

Type	Expect
Het/Hom SNP (at least 2bp from another small variant)	>3M
Het/Hom Insertion/Deletion, Length Polymorphism	~500K
Het/Hom Substitutions, Length Conserving and Length Altering	~75K
Complex Variants	~25K
Partial Information (haploid calls and/or N's in assembly)	~100K

Very rough typical call-rate numbers for germ-line DNAs of causasian or asian decent.

14



Humans Are Not a List of SNPs

Complete genomics

```
 Position:  123 456 789     Protein          Event
Reference:  GTA CGT GGC     Val Arg Gly
Allele 1:   GTA CGT GGC     Val Arg Gly      (reference)
Allele 2:   GTA TGA GGC     Val STOP         (nonsense)
```

Three nucleotide heterozygous substitution as called by local *de novo* assembly

```
Reference:  GTA CGT GGC     Val Arg Gly
Het SNP 1:        A         Val Arg Gly      (synonymous)
Het SNP 2:      T           Val Cys Gly      (non-synonymous)
```

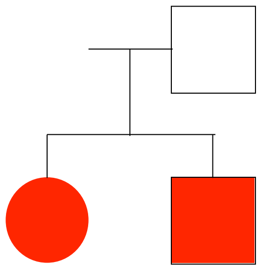
Locus re-coded as two heterozygous SNPs with loss of phase information

- There are various complexities if attempting to call humans as a list of SNPs…
  - Recoding is robust when SNPs are well separated and alignments of alleles against reference are unambiguous. Recoding is not robust when these are not so.
  - Variant alleles from *de novo* assembly can have different lengths, and both alleles can be different lengths than the corresponding reference sequence. Recoding can be hard to define consistently in such cases.
  - One must always remember phase!

anonymized CG customer data 15

Complete Genome Sequences of a Family

2 Parents + 2 Children



Children Affected
By Two Separate
Mendelian Diseases:

- Miller Syndrome
- Ciliary Diskinesia

Goals of Study

- Determine cause of Mendelian diseases affecting both children
- Measure *de novo* mutation rate in children
  ~ 1.1 x 10^{-8}
- Develop analysis methods for future studies
- Benchmark performance of genome sequencing platform
  - Comparison to independent exome data
  - Large validation data set from de novo mutation study
  - Consider the 25% of the genome identical between the two children as a reproducibility study

Roach et al. Science 2010; Roach et al. AJHG 2011

Potential Causative Variants Discovered in Family of Four

Strategy:

- Assume recessive inheritance of novel loss-of-function mutations. Allow for simple recessive or compound-heterozygous LOF mutations affecting a single gene/element. Also tested a dominant model.
- Assume causal homogeneity for the affected children: Restrict analysis to regions of the genome with identical DNA from mother and father (22%) in both, leveraging the fine scale recombinational map.
- Disregard mendelian inconsistent sites, leveraging error detection possible in family with fine structure recombination map.

Results:

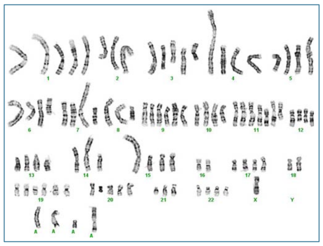
- Only **nine** candidate causative loci in annotated genome regions fitting recessive or compound-heterozygous genetic model:
  - Four protein-coding changes:
    - **DHODH, DNAH5**, KIAA0556, CES1
  - One Intronic, near splice site
  - One in UTR, putative signal sequence
  - Four in non-protein coding RNA genes

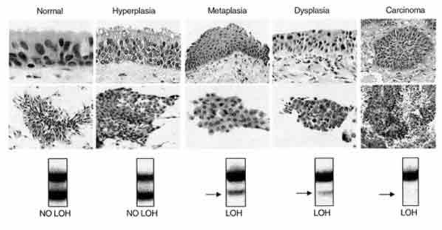
DHODH is the cause of Miller Syndrome and DNAH5 is the cause of Ciliary Diskinesis in the two children.

Roach et al. Science 2010; Ng et al. Nature 2009

Tumor Sample can be both Aneuploid and Heterogeneous

- Heterogeneity can arise due to:
  - Normal/Stromal tissue contamination within tumor sample
  - Multiple tumor populations within tumor sample

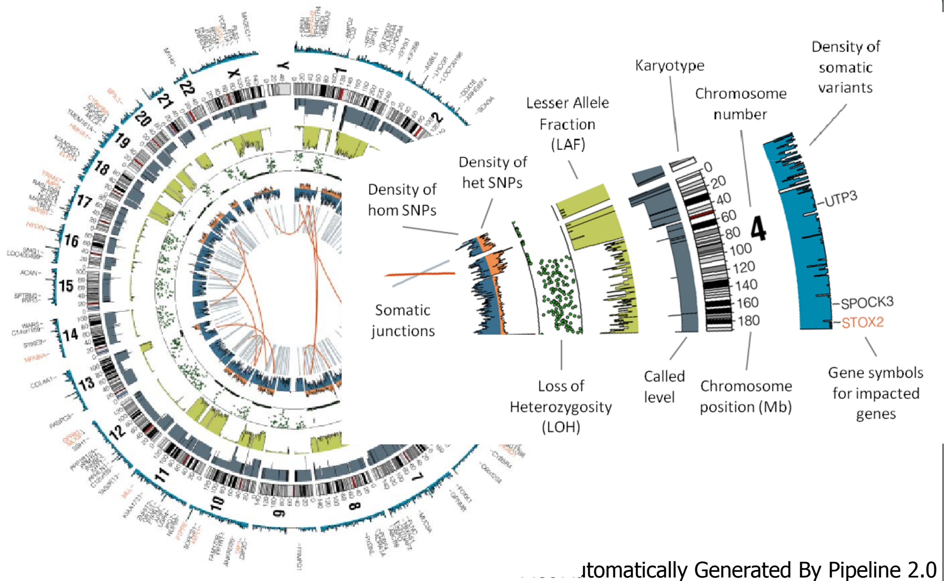




http://www.bentham.org/cmm/sample/cmm1-1/miatra/Miatra-fig3-pg159.jpg

- Aneuploidy means that copy numbers can vary substantially
  - Baseline or mean/median for sample is not diploid (CN=2)
  - Given heterogeneity, copy numbers may not be integers

Comprehensive Assessment of Somatic Variation in Tumor-Normal Pairs:



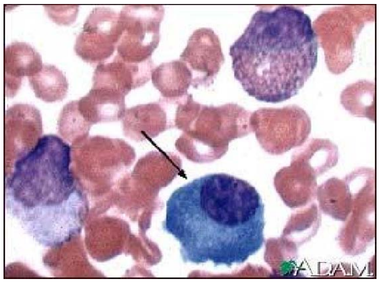
utomatically Generated By Pipeline 2.0

Circos Software: Krzywinski, et al. Genome Research 2009.

Waldenstrom's Macroglobulinemia: Consistent Activating Mutation

DANA-FARBER CANCER INSTITUTE

- Sequenced 10 matched tumor-normal pairs
  - Older CG Pipeline 1.10; Standard-Depth Sequencing (~55x average)
- Single specific point mutation in MYD88 found in 90% of TN pairs
  - One T/N pair missing the somatic SNP call had it in 12% of the reads
  - This specimen had significant heterogeneity, according to pathologist
- Gain-of-function: Variant constitutively activates IRAK and NF-kB
  - Validation and downstream functional studies started within weeks of receiving genome sequences
- Credits:
  - Steve Treon MD PhD
  - Zachary Hunter PhD
  - et al.
- Presented at ASH 2011 Meeting
  - Manuscript in press

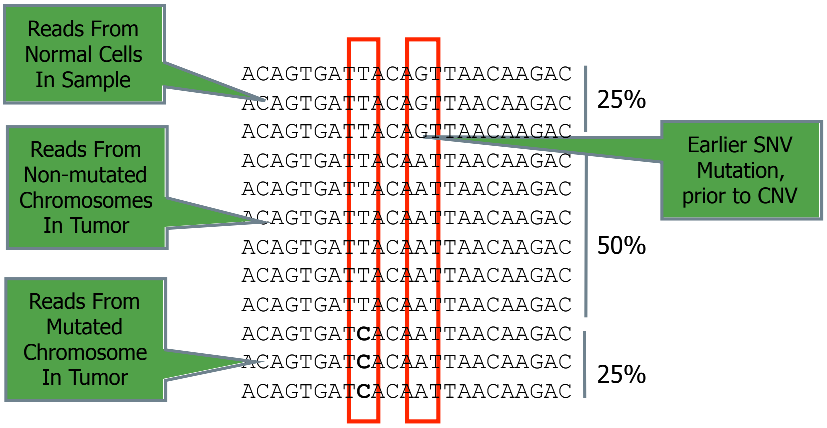


Effect of Heterogeneity and Aneuploidy on Small Variant Detection

Complete genomics

Small variants may be present in only a small fraction of the reads…

Consider a tumor sample with 25% normal contamination and a ploidy 3 CNV region with a somatic mutation in the minor allele. One would expect…



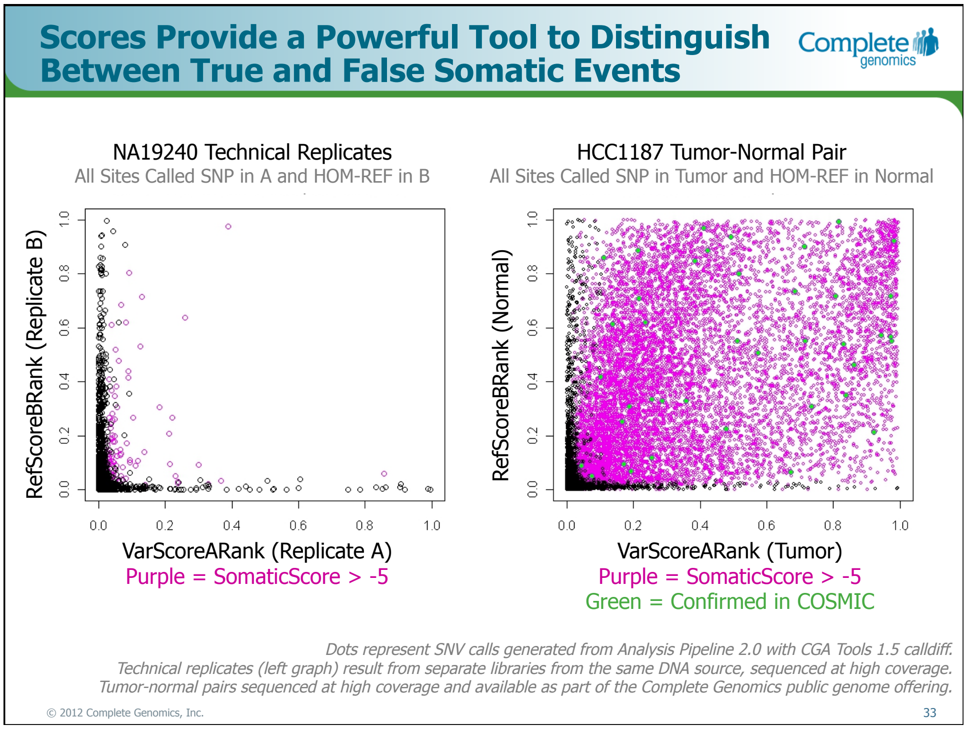
Reads From Normal Cells In Sample

Reads From Non-mutated Chromosomes In Tumor

Reads From Mutated Chromosome In Tumor

Earlier SNV Mutation, prior to CNV

```
ACAGTGATTACAGTTAACAAGAC
ACAGTGATTACAGTTAACAAGAC   25%
ACAGTGATTACAGTTAACAAGAC
ACAGTGATTACAATTAACAAGAC
ACAGTGATTACAATTAACAAGAC
ACAGTGATTACAATTAACAAGAC
ACAGTGATTACAATTAACAAGAC   50%
ACAGTGATTACAATTAACAAGAC
ACAGTGATTACAATTAACAAGAC
ACAGTGATCACAATTAACAAGAC
ACAGTGATCACAATTAACAAGAC   25%
ACAGTGATCACAATTAACAAGAC
```



Scores Provide a Powerful Tool to Distinguish Between True and False Somatic Events

NA19240 Technical Replicates

All Sites Called SNP in A and HOM-REF in B

HCC1187 Tumor-Normal Pair

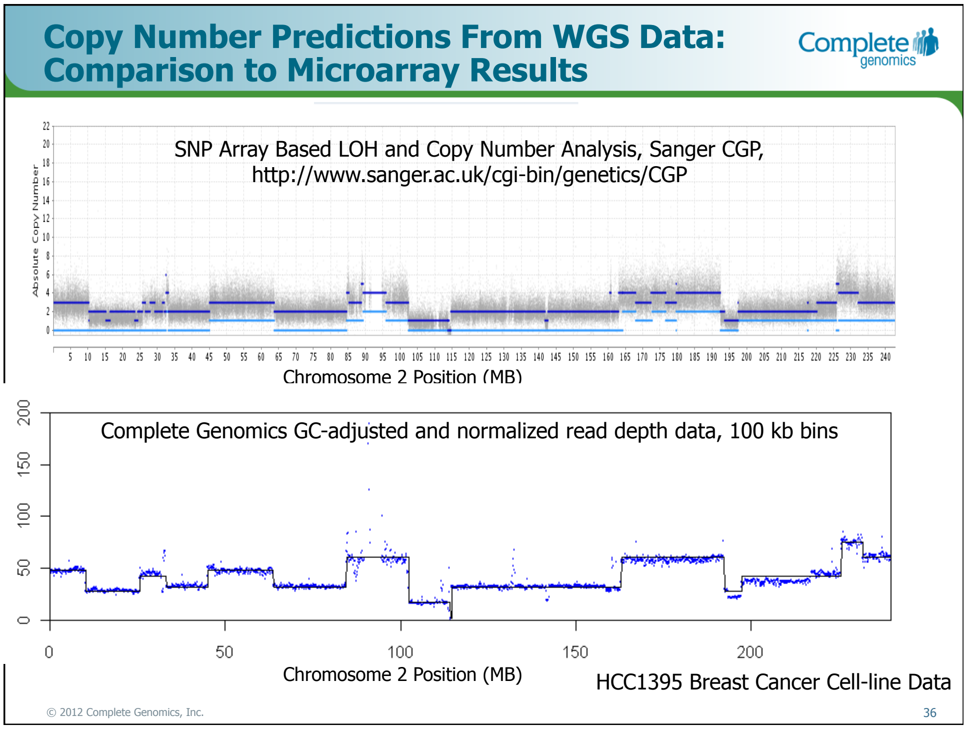
All Sites Called SNP in Tumor and HOM-REF in Normal

Purple = SomaticScore > -5

Purple = SomaticScore > -5

Green = Confirmed in COSMIC

Dots represent SNV calls generated from Analysis Pipeline 2.0 with CGA Tools 1.5 calldiff.
Technical replicates (left graph) result from separate libraries from the same DNA source, sequenced at high coverage.
Tumor-normal pairs sequenced at high coverage and available as part of the Complete Genomics public genome offering.

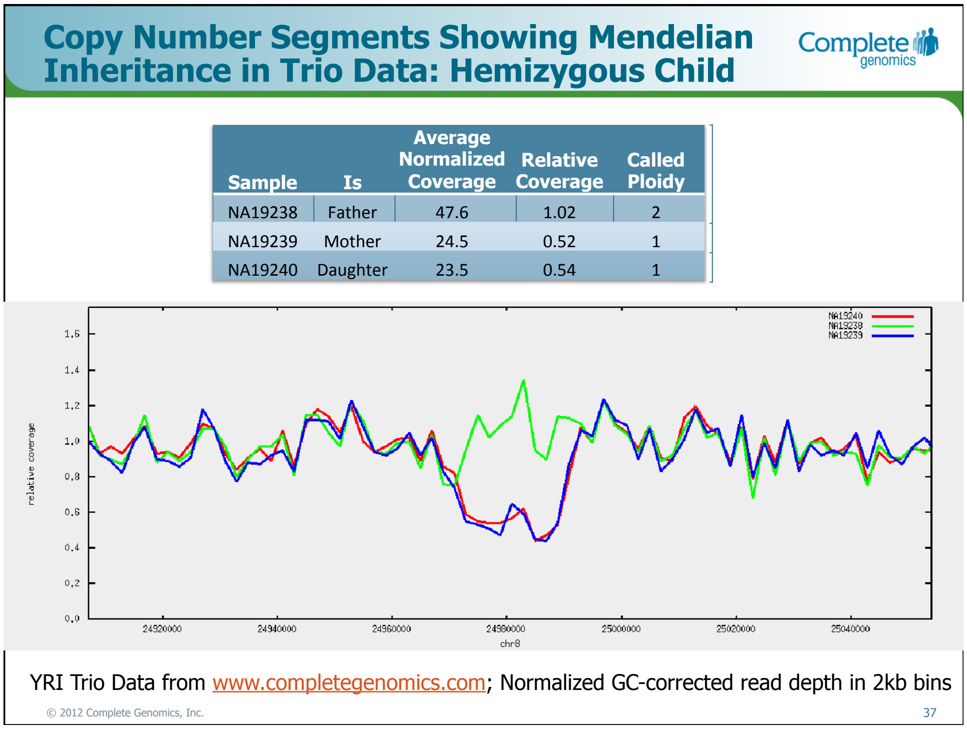


Copy Number Predictions From WGS Data: Comparison to Microarray Results

SNP Array Based LOH and Copy Number Analysis, Sanger CGP,
http://www.sanger.ac.uk/cgi-bin/genetics/CGP

Complete Genomics GC-adjusted and normalized read depth data, 100 kb bins

HCC1395 Breast Cancer Cell-line Data



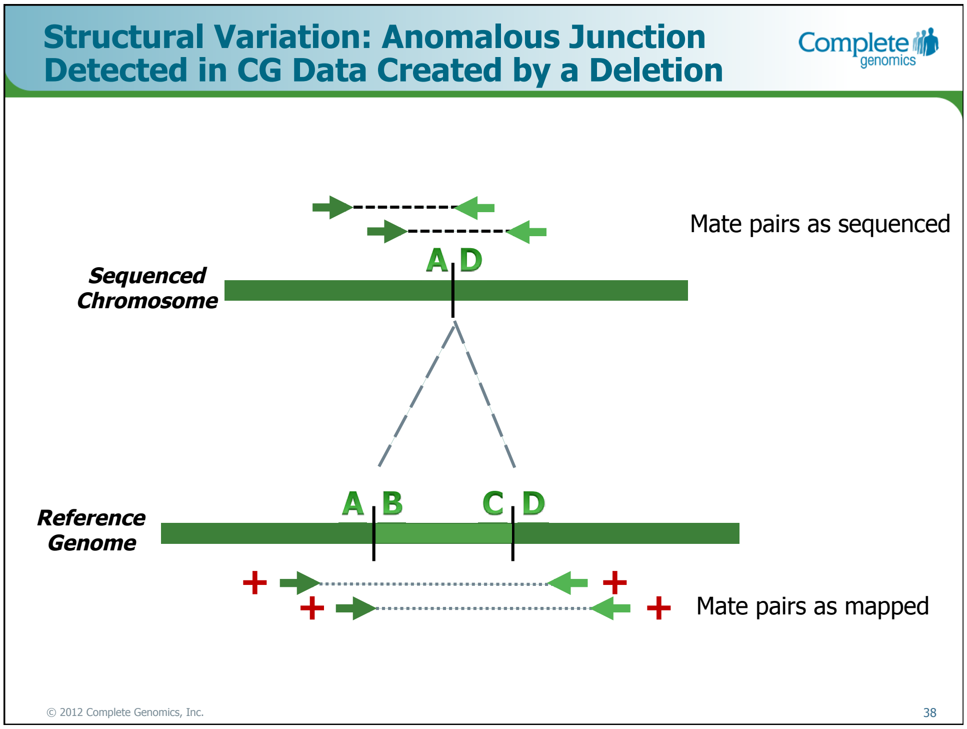
Copy Number Segments Showing Mendelian Inheritance in Trio Data: Hemizygous Child

Complete genomics

Sample	Is	Average Normalized Coverage	Relative Coverage	Called Ploidy
NA19238	Father	47.6	1.02	2
NA19239	Mother	24.5	0.52	1
NA19240	Daughter	23.5	0.54	1

YRI Trio Data from www.completegenomics.com; Normalized GC-corrected read depth in 2kb bins

37



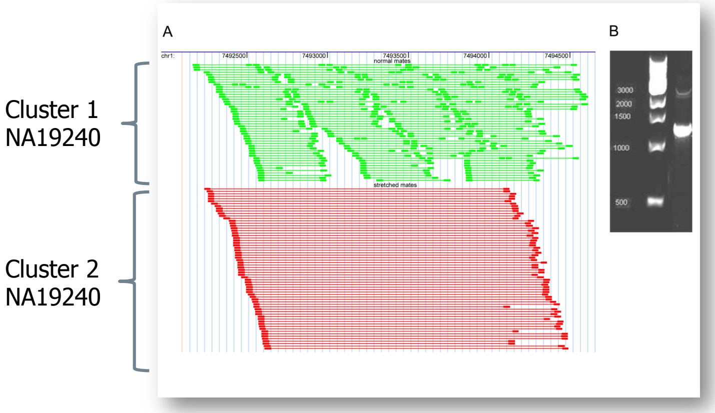
Structural Variation: Anomalous Junction Detected in CG Data Created by a Deletion

Complete genomics

38

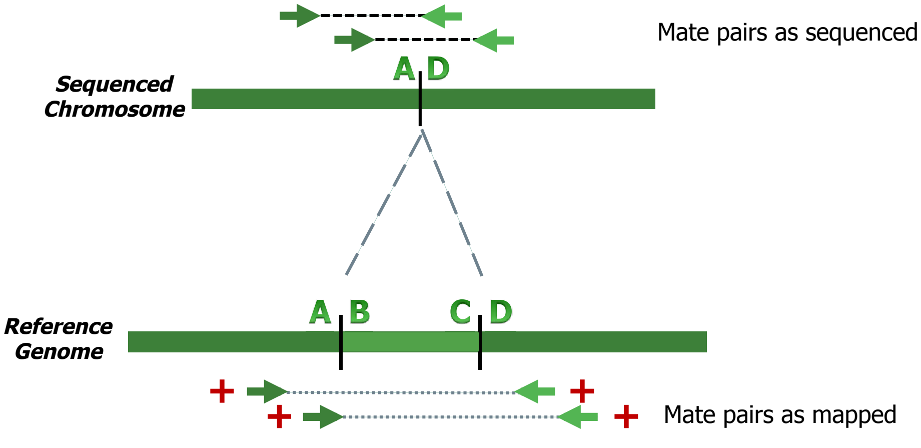
Read-Pair Analysis can Identify Structural Variations in CG Data

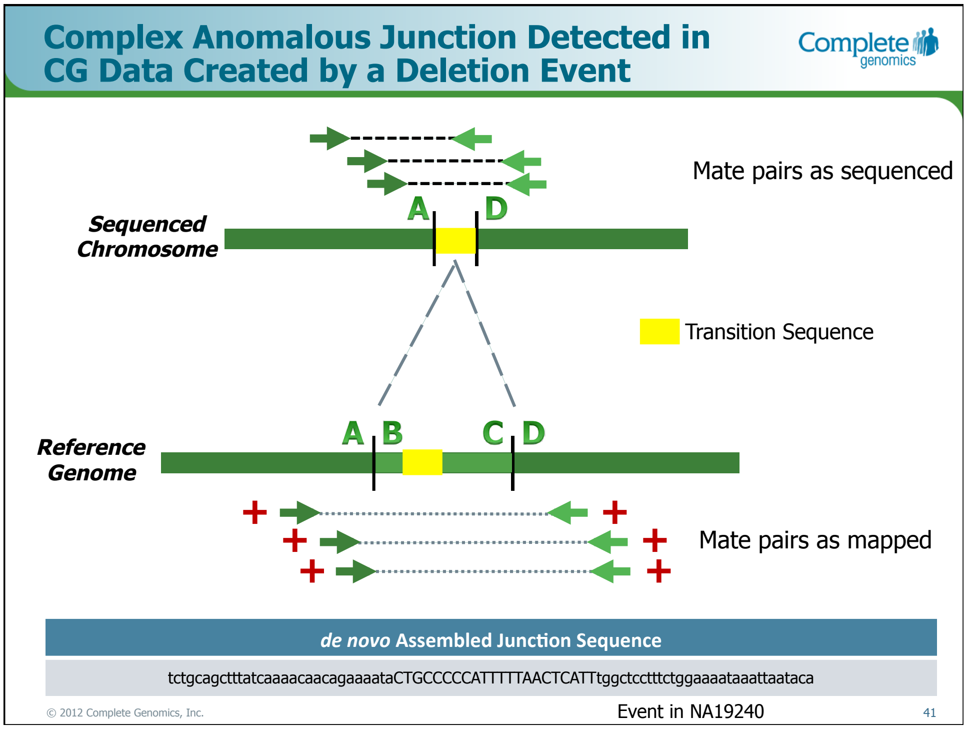
Example: Two distinct groups of clones were identified in one individual in this 1,500bp region of chromosome 1. Data show heterozygous deletion of an Alu element validated by PCR.



Drmanac et al. Science 2010

Structural Variation: Anomalous Junction Detected in CG Data Created by a Deletion



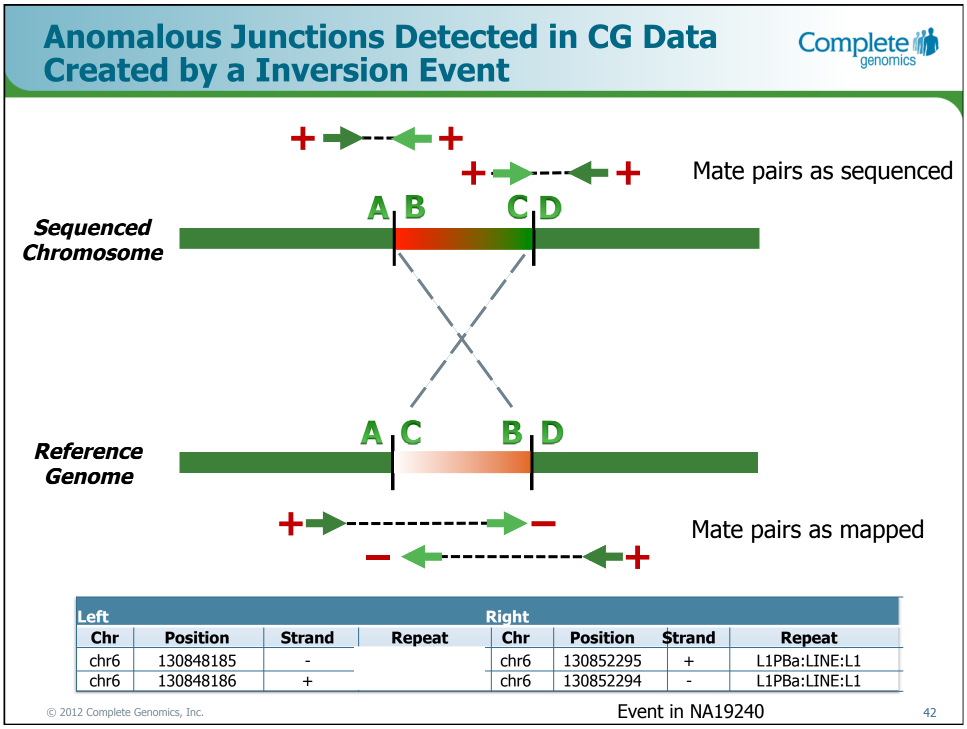


Complex Anomalous Junction Detected in CG Data Created by a Deletion Event

de novo Assembled Junction Sequence

tctgcagctttatcaaaacaacagaaaataCTGCCCCCATTTTTAACTCATTtggctcctttctggaaaataaattaataca

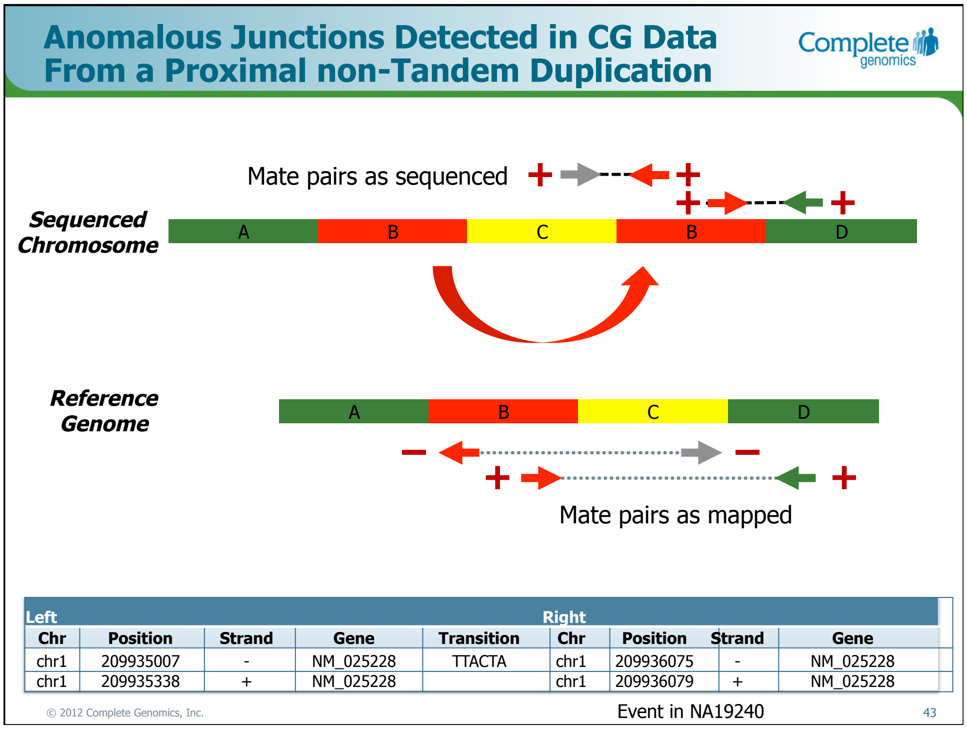
Event in NA19240



Anomalous Junctions Detected in CG Data Created by a Inversion Event

Left				Right			
Chr	Position	Strand	Repeat	Chr	Position	Strand	Repeat
chr6	130848185	-		chr6	130852295	+	L1PBa:LINE:L1
chr6	130848186	+		chr6	130852294	-	L1PBa:LINE:L1

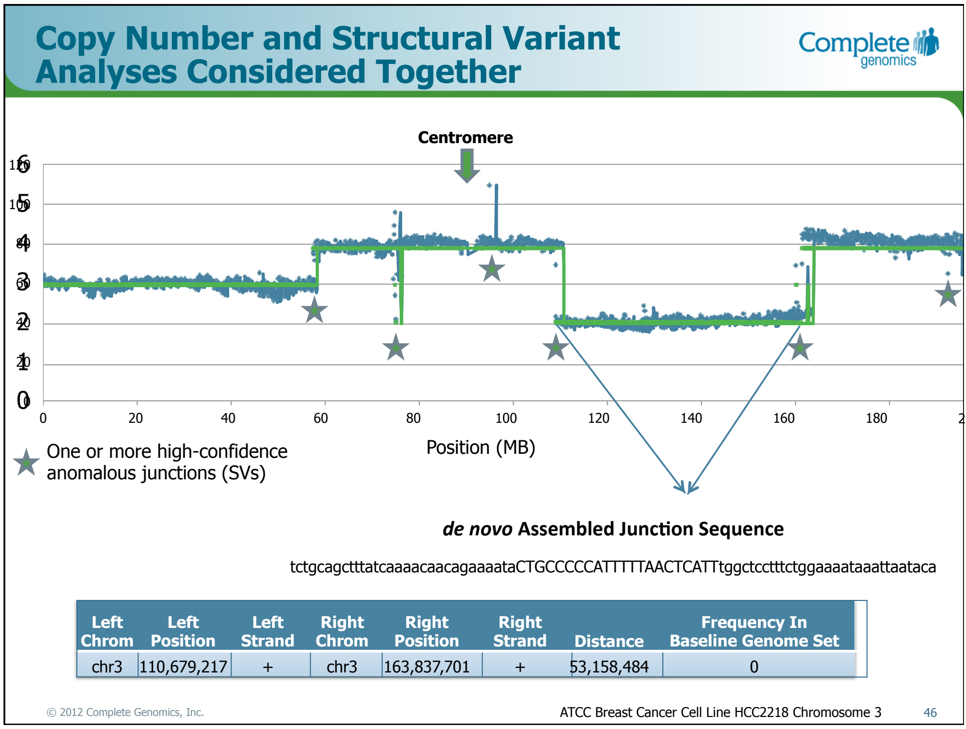
Event in NA19240



Anomalous Junctions Detected in CG Data From a Proximal non-Tandem Duplication

Left					Right			
Chr	Position	Strand	Gene	Transition	Chr	Position	Strand	Gene
chr1	209935007	-	NM_025228	TTACTA	chr1	209936075	-	NM_025228
chr1	209935338	+	NM_025228		chr1	209936079	+	NM_025228

Event in NA19240



Copy Number and Structural Variant Analyses Considered Together

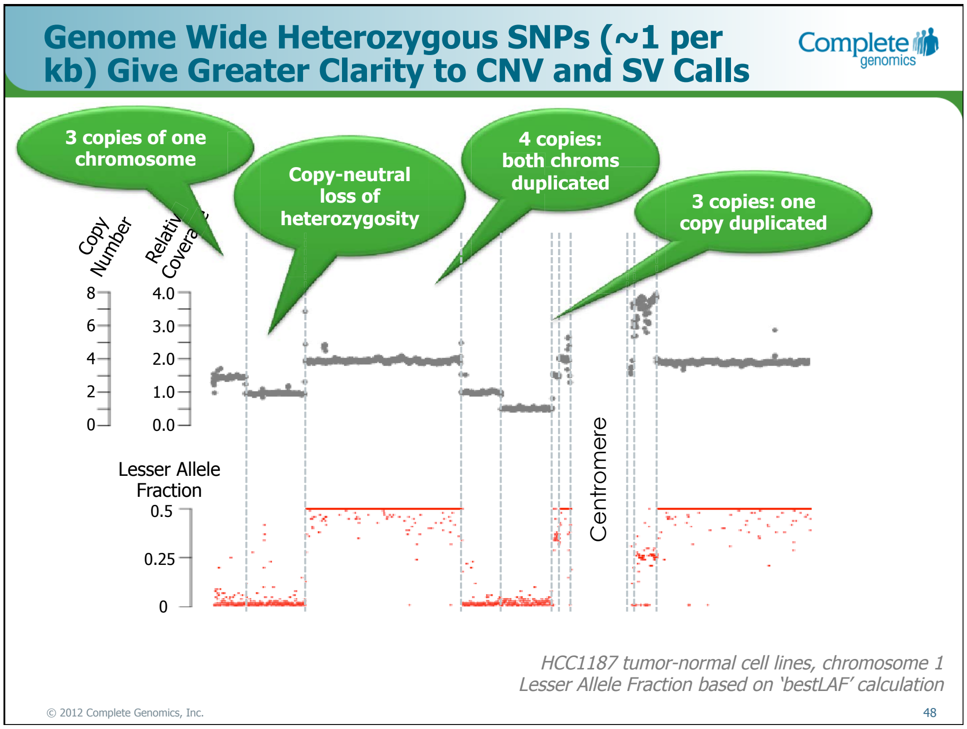
★ One or more high-confidence anomalous junctions (SVs)

***de novo* Assembled Junction Sequence**

tctgcagctttatcaaaacaacagaaaataCTGCCCCCATTTTTAACTCATTtggctcctttctggaaaataaattaataca

Left Chrom	Left Position	Left Strand	Right Chrom	Right Position	Right Strand	Distance	Frequency In Baseline Genome Set
chr3	110,679,217	+	chr3	163,837,701	+	53,158,484	0

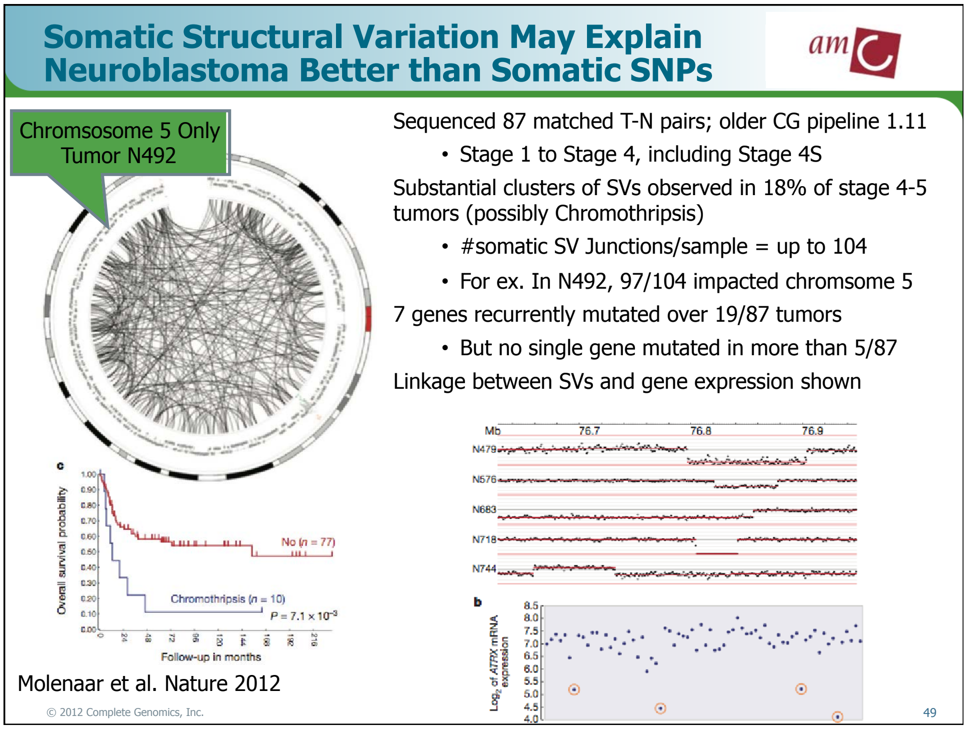
ATCC Breast Cancer Cell Line HCC2218 Chromosome 3



Genome Wide Heterozygous SNPs (~1 per kb) Give Greater Clarity to CNV and SV Calls

Complete genomics

HCC1187 tumor-normal cell lines, chromosome 1
Lesser Allele Fraction based on 'bestLAF' calculation



Somatic Structural Variation May Explain Neuroblastoma Better than Somatic SNPs

Sequenced 87 matched T-N pairs; older CG pipeline 1.11

- Stage 1 to Stage 4, including Stage 4S

Substantial clusters of SVs observed in 18% of stage 4-5 tumors (possibly Chromothripsis)

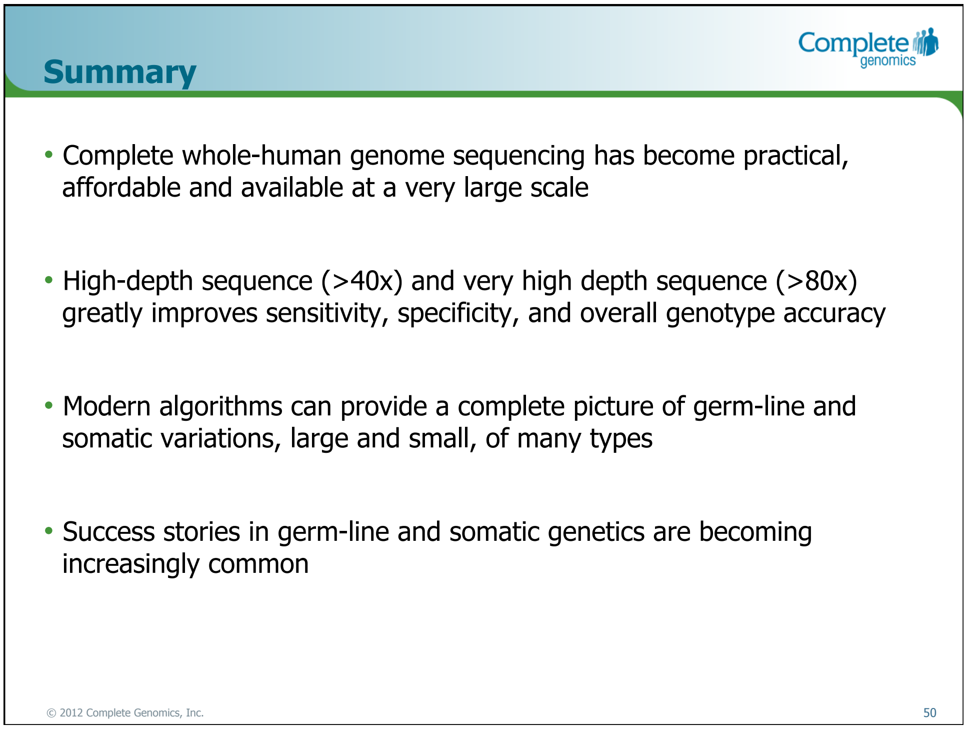
- #somatic SV Junctions/sample = up to 104
- For ex. In N492, 97/104 impacted chromsome 5

7 genes recurrently mutated over 19/87 tumors

- But no single gene mutated in more than 5/87

Linkage between SVs and gene expression shown

Molenaar et al. Nature 2012



Summary

- Complete whole-human genome sequencing has become practical, affordable and available at a very large scale
- High-depth sequence (>40x) and very high depth sequence (>80x) greatly improves sensitivity, specificity, and overall genotype accuracy
- Modern algorithms can provide a complete picture of germ-line and somatic variations, large and small, of many types
- Success stories in germ-line and somatic genetics are becoming increasingly common