Complete genomics

Analysis of Thousands of Whole Human Genome Sequences

Steve Lincoln
BioC Seattle, July 28, 2011

Complete Genomics: Whole Human Genome Sequencing as a Service

Complete genomics

Some Selected End-Users

BRIGHAM AND WOMEN'S HOSPITAL
A Teaching Affiliate of Harvard Medical School
UT SOUTHWESTERN MEDICAL CENTER
Erasmus MC
BROAD INSTITUTE
Pfizer
CANCER RESEARCH UK
RIKEN
THE UNIVERSITY of TEXAS HEALTH SCIENCE CENTER AT HOUSTON
VIB
am
Lilly
INSTITUTE FOR Systems Biology
NATIONAL CANCER INSTITUTE
ICR The Institute of Cancer Research
STANFORD UNIVERSITY
The Children's Hospital of Philadelphia®

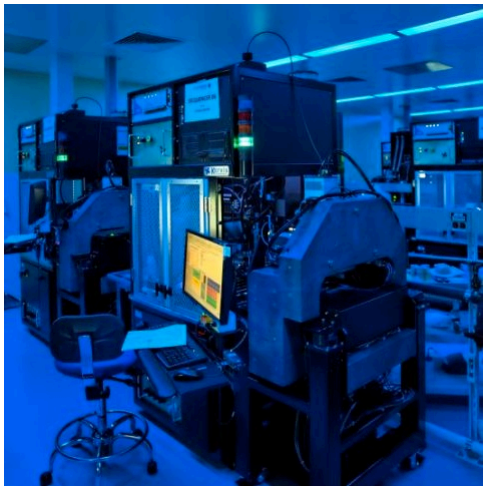
Our Service

- Service lab providing DNA sequencing and data analysis
- High-depth complete human genome sequences
- Use a specialized and custom-made sequencing platform

Production To Date

- CG V3 instruments: ~1 human genome per day each at ~60x
- Over 1400 genomes delivered to customers (March 2011)
- Capacity over 450 human genomes per month (June 2011)
- Turn around: quote 90-120 days, current median ~68 days

Complete Genomics Custom Human Genome Sequencing Instruments



Attributes

- Open architecture for maintainability and rapid and frequent upgrades
- High Throughput: ~1 Human Genome per day per Instrument currently
- Substantial upgrades in development for 2011 and beyond

High Coverage of Sequenced Human Genomes

Metric	Non-Tumor Genomes	Tumor Genomes
Average Gross Mapped Genome Coverage	>55X	> 55X
Average Genome Read Coverage ≥ 10X	98.27%	98.26%
Average Unique Genome Read Coverage ≥ 10X	96.05%	95.98%
Average Exome Read Coverage ≥ 10X	96.94%	96.93%
Average Unique Exome Read Coverage ≥ 10X	94.42%	94.42%

Gross mapping includes both single and double-end placements.
Read coverage requires consistent paired-end placement(s) weighted by mapping likelihoods.
Unique coverage requires a single consistent paired-end placement preferred over any other by 100:1.

Measurements against the complete ~2.85 GB NCBI reference genome.
Results are prior to local *de novo* assembly.

Data from previous 90 days as of March 29, 2011

Applications of Complete Whole Genome Sequencing

Somatic Mutations in Lung Cancer (Genentech)

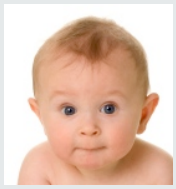
- Compared Resected Primary Tumor to matched Normal
- ~50,000 Somatic SNPs at >90% validation rate
- 79 Somatic Structural Variations at a 66% validation rate
- **Finding: 1 Point Mutation per 3 Cigarettes smoked**

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: 1.1×10^{-8}
- Further benchmarked accuracy of the CGA™ platform

Roach et al., Science 2010

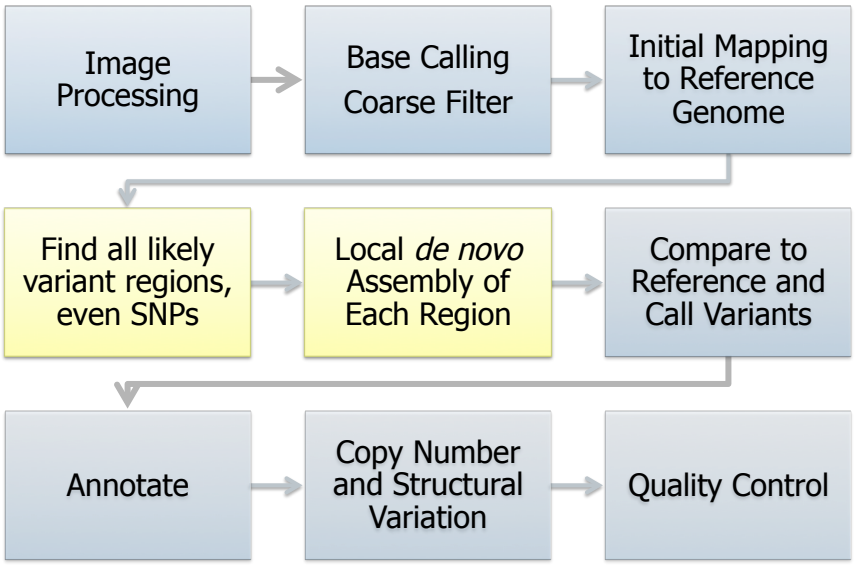


Affected Individual with Extreme Phenotype (UTSW)

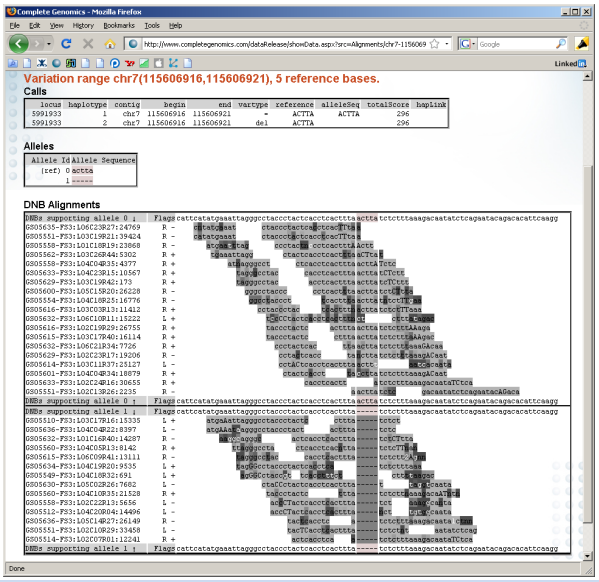
- 11-Month Old with Severe Hypercholesterolemia
- Blood Test and Traditional DNA tests failed to identify cause
- **Genome sequencing showed required protein absent which had been missed by other genetic tests**

Rios et al., HMG 2010

Complete Genomics Uses a Two Step Mapping and Assembly Process



Example of Diploid Local *de novo* Assembly: Heterozygous 5bp Deletion in One Individual



Accuracy of Complete Genomics Data: 1 Error in 300,000 Bases on Single Samples

Consensus Error Rate Single Sample Basis	• 1 in 125,000 coding; 1 in 91,000 genome-wide (2009 data) • Errors include False+ and False- (false hom-ref calls) • Recent data at same customer: 1 in 300,000 genome-wide *Roach et al., Science 2010; and Institute for Systems Biology (unpublished)*
Comparative Analysis Accuracy	• Mendelian Inheritance errors: 1 per 300,000 bp (2009 data) • Yoruban trio child errors: 1 per 420,000 bp (2010 data) • T-N pairs: <1 false+ somatic SNV per 1-5 MB *Roach et al., Science 2010; YRI trio data on www.completegenomics.com*
Novel SNV False Discovery Rate	• ~0.4% on a single sample basis (late 2009 data) • ~0.2% in replicate sequences (2010 data) • This is consistent with above error rates *Dramanac et al., Science 2010*
HapMap Concordance	• 99.91% concordance with HapMap II Infinium Subset • 99.97% concordance allowing zygosity differences *YRI Trio Data. www.completegenomics.com Jan 2011*
CNV/SV Calls	• 78% SV Sanger Validation Rate (higher in some tumors) • 96% of CG calls overlap CNVEs in Conrad et al. (Nature 2010) • 95% CNV reproducibility *NA19240 Data. www.completegenomics.com Jan 2011*

Genome-wide mappings vs. Local de novo assemblies of variant regions

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

Example:

```
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
- Allele 2: TCG to CCCTC length-altering block substitution
- Locus (yellow box) is called "complex" in CG masterVar file

Type	%loci	Expect
Het/Hom SNP (at least 2bp from another small variant)	84.1%	~3M+
Het/Hom Insertion/Deletion, Length Polymorphism	8.5%	~400K
Het/Hom Substitutions, Length Conserving and Length Altering	1.8%	~65K
Complex Variants	0.7%	~25K
Partial Information (haploid calls and/or N's in assembly)	4.9%	~150K

NA19240 Pipeline v1.8 chr21. Typical call-rate numbers for Caucasian germ-line DNAs.

Can complex variants be recoded as SNPs? Well, not so much…

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 and issues in attempting to do so…
  - 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.
  - Alleles from de novo assembly can have different lengths, and both different that the corresponding reference.
  - One must always remember phase!

Validated non-coding variants (small and large) in various human diseases

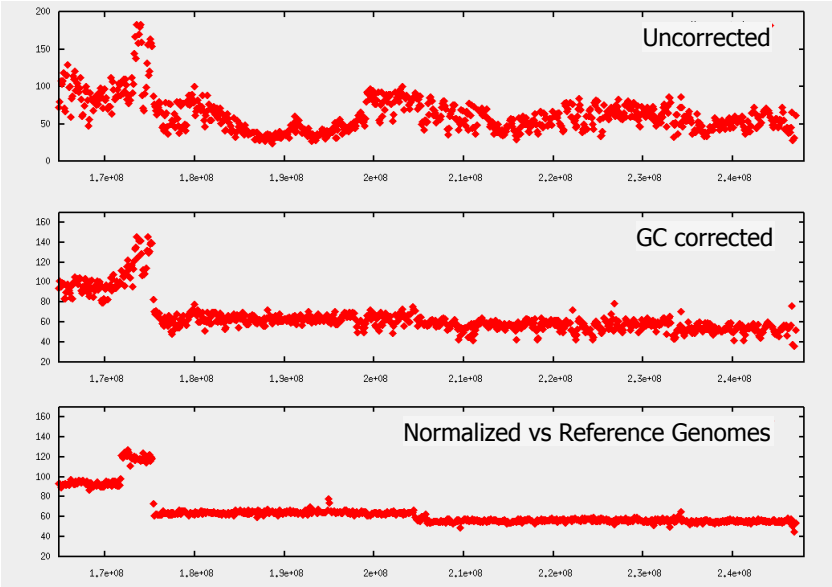
Mutations in:

- ✓Promoters
- ✓UTR regulatory regions
- ✓Intronic splicing regulators
- ✓Genomic regulatory regions (for ex. enhancers)
- ✓Non-coding RNAs
- ✓Copy number variants in and near genes
- ✓Copy-neutral structural variants in and near genes

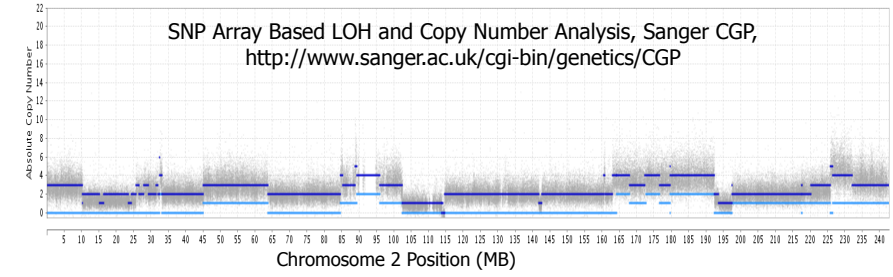
Disease Area

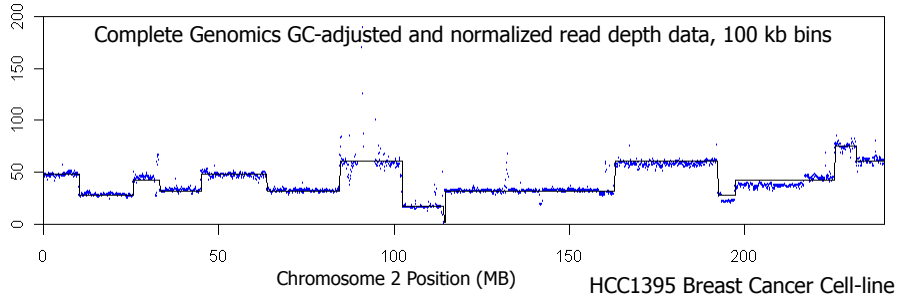
- –Allergies and Asthma
- –Hypertension
- –Coronary heart disease
- –Beta Thalassemia
- –Developmental disorders
- –HIV Susceptibility
- –Psycoaffective disorders
- –Alzheimer's disease
- –Many Cancers

Copy Number Variation: GC correction and baseline normalization



Copy Number Predictions From CG Data: Comparison to Microarray Results

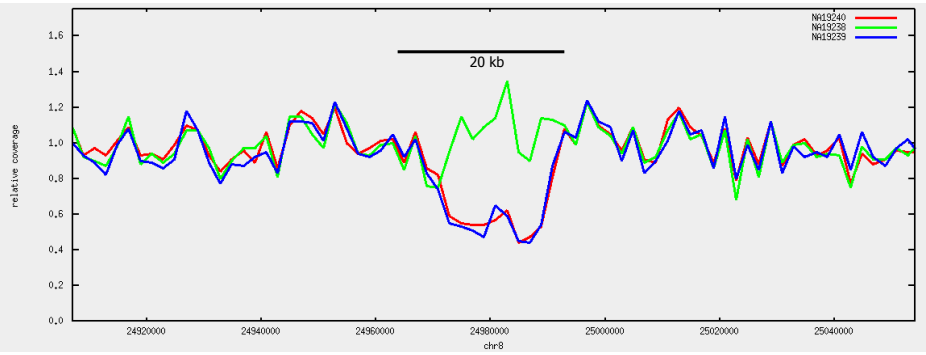




HCC1395 Breast Cancer Cell-line Data

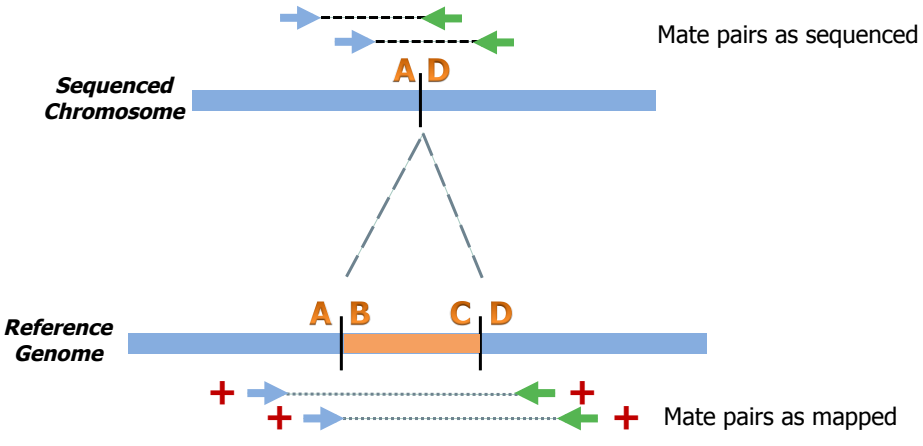
Copy Number Segments Showing Mendelian Inheritance in Trio Data: Hemizygous Child

Sample	Is	chr	begin	end	Average Normalized Coverage	Relative Coverage	Called Ploidy	Known CNV?
NA19238	Father	chr8	23509854	31363854	47.6	1.02	2	
NA19239	Mother	chr8	24971854	24989854	24.5	0.52	1	dgv.1:Variation_1191
NA19240	Daughter	chr8	24971854	24989854	23.5	0.54	1	dgv.1:Variation_1191

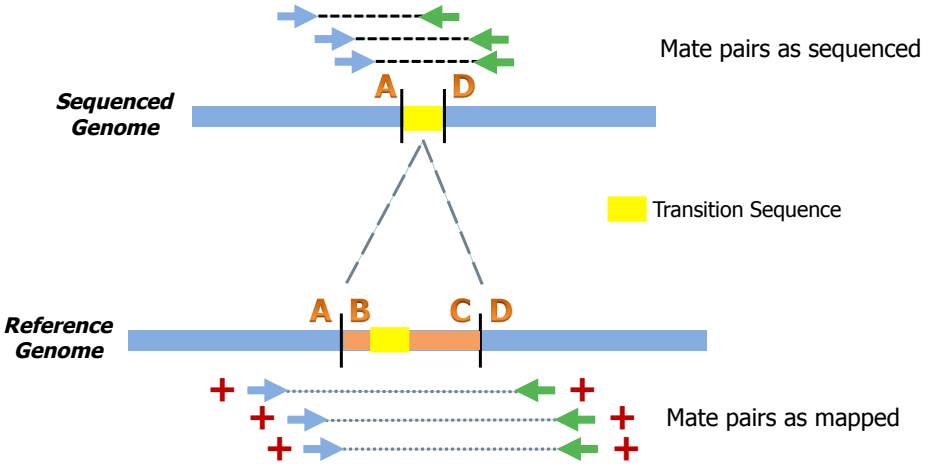


YRI Trio Data from www.completegenomics.com; Normalized GC-corrected read depth in 2kb bins

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



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



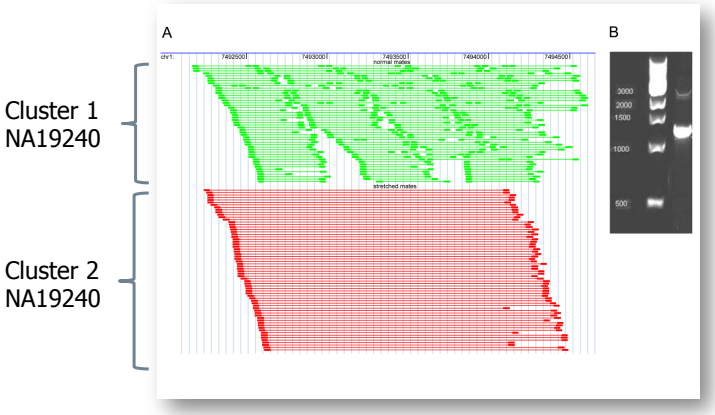
***de novo* Assembled Junction Sequence**

tctgcagctttatcaaaacaacagaaaataCTGCCCCCATTTTTAACTCATTtggctcctttctggaaaataaattaataca

Event in NA19240

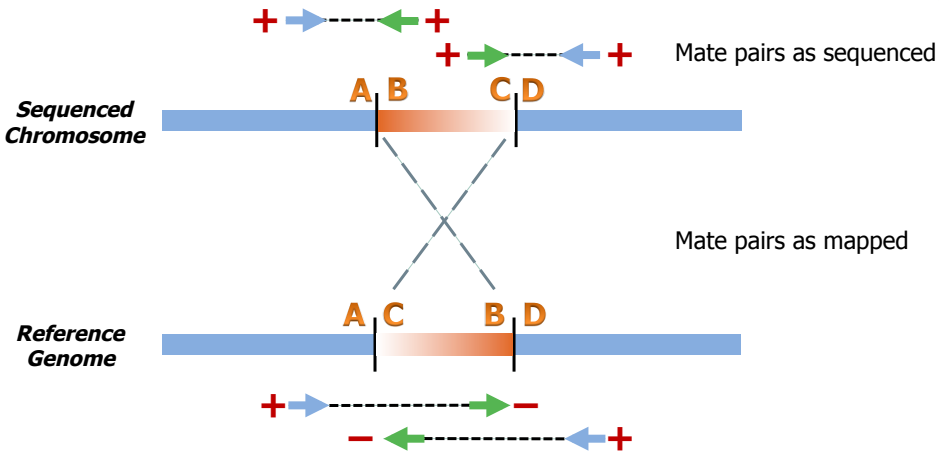
Read-Pair Analysis can Identify Structural Variations in CG Data

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

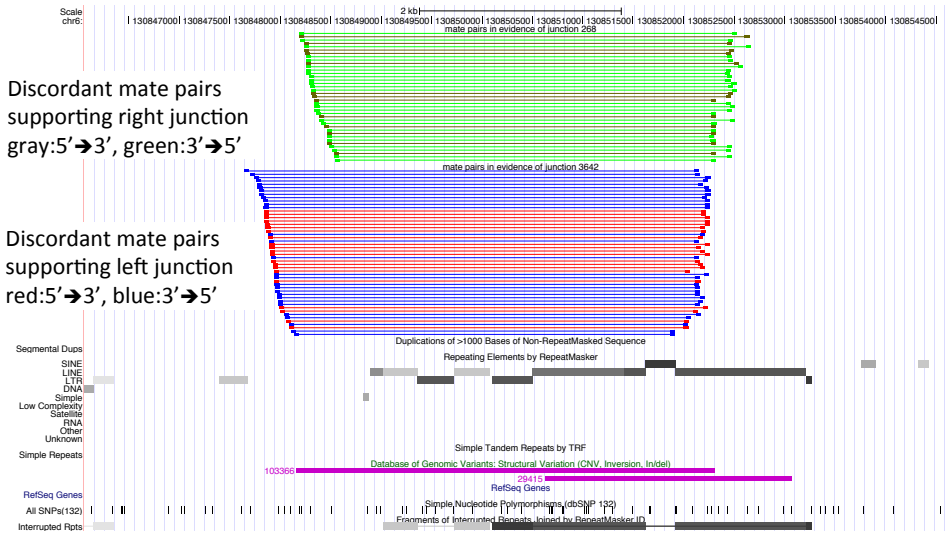
Anomalous Junctions Detected in CG Data Created by a Inversion Event



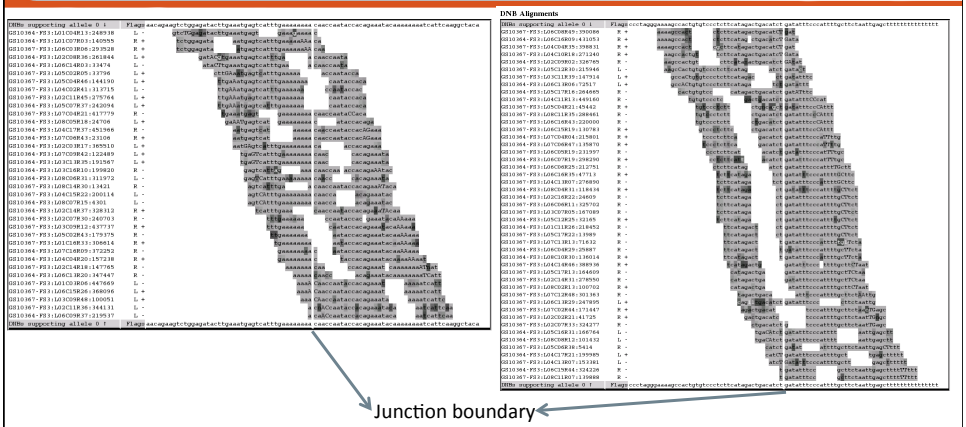
Left				Right			
Chr	Position	Strand	Repeat	Chr	Position	Strand	Repeat
chr6	130848185	-		chr6	130852295	+	L1PBa:LINE:L1
chr6	130848186	+		chr6	130852294	-	L1PBa:LINE:L1

Event in NA19240

Mate evidence



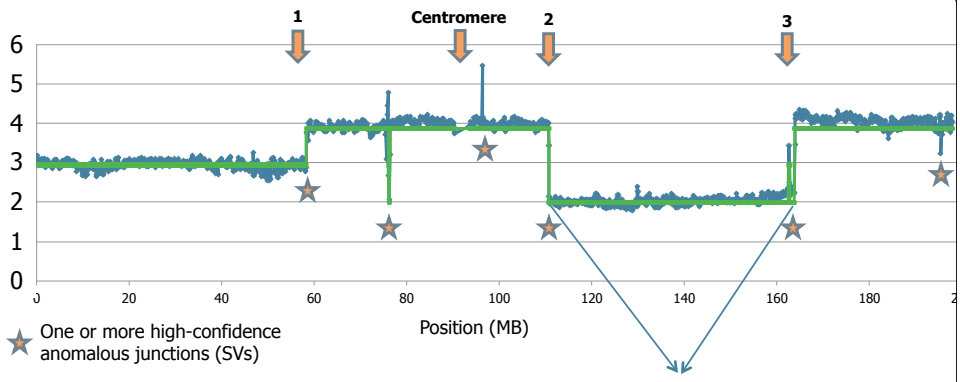
Sequence evidence for left and right junctions



Left-junction
Sequence to the left: +strand outside the inversion
Sequence to the right: -strand at the end of the inverted segment (from the perspective of the reference genome)

Right-junction
Sequence to the right: +strand, outside the inversion
Sequence to the right: -strand at the start of the inverted segment (from the perspective of reference genome)

Copy Number and Structural Variant Analyses Considered Together



***de novo* Assembled Junction Sequence**

tctgcagctttatcaaaacaacagaaaataCTGCCCCCATTTTTAACTCATTtggctccttctggaaaataaattaataca

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 Breat Cancer Cell Line HCC2218 Chromosome 3 – Copy Number Using 100KB Windows

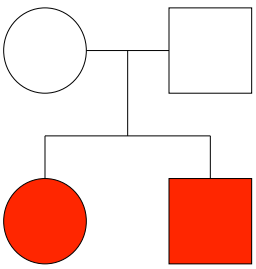
Complete genomics

Using Whole Human Genome Sequences

Institute for Systems Biology, Seattle Complete Genome Sequences of a Family

INSTITUTE FOR Systems Biology

2 Parents + 2 Children



Four Genomes Sequenced by Complete Genomics Children independently exome sequenced

Roach et al. Science 2010

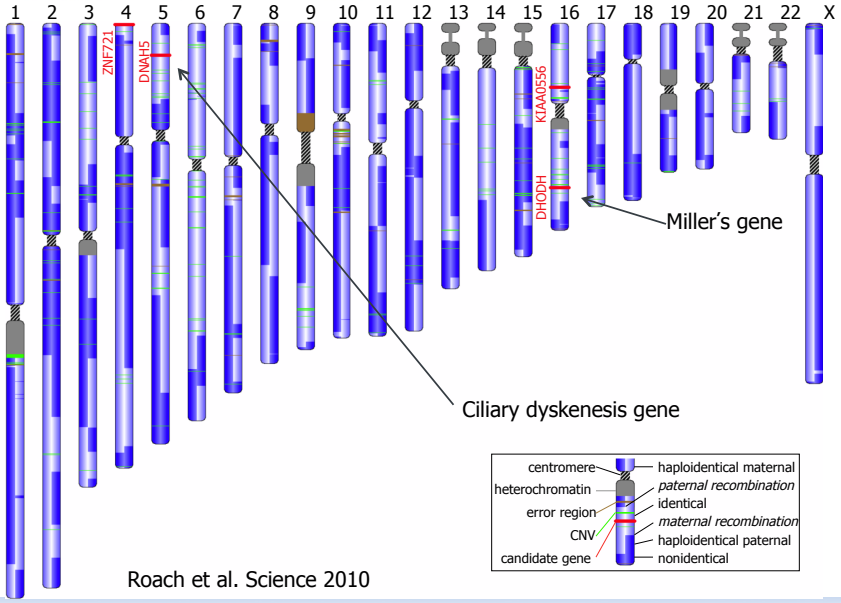
Multiple Data Accuracy Analyses Possible

- Mendelian Inconsistencies
- Compare CG genome sequence to independent exome sequence
- Compare CG genome sequence to targeted resequencing and genotyping results
- Consider as replicates ~25% of genome where both children are identical twins

Error rate estimates for Complete Genomics data (in called bases):

- In Exome: 8.1×10^{-6}
- Genome-wide: 1.1×10^{-5}
- Family False+: 3.3×10^{-6}

Genome Sequences allows for construction of a fine structure recombination map



Roach et al. Science 2010

Potential Causative Variants Discovered in Miller Syndrome Family

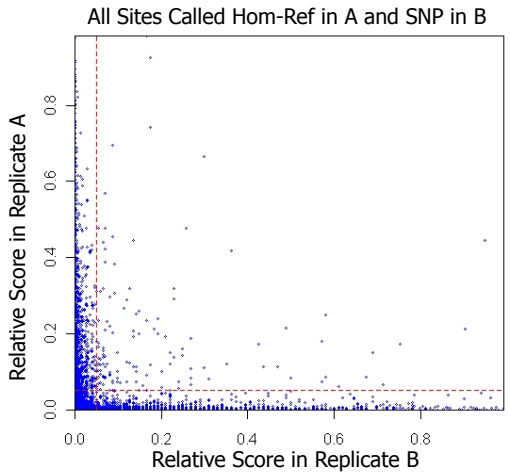
Strategy:

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

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

Evidence Both For and Against Variants Increases Power to Detect Somatic Variants



Complete Data

Novel False "Somatic SNPs"
= 4,258 (Development Pipeline)

Each Such Error Must Be Either:
- False Positive in Replicate B, or
- False Negative in Replicate A

Inside Red Lines

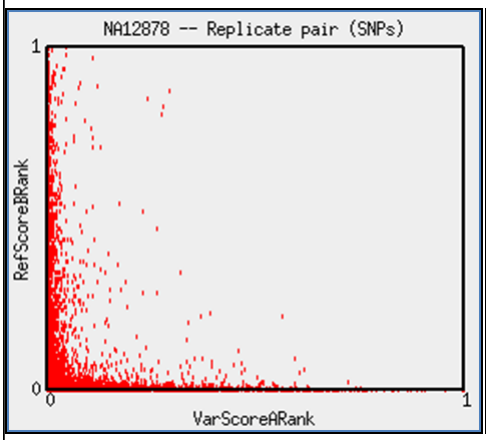
Weakest "Somatic SNPs" Removed
= 5% of True+ (concordant calls)
= 91% of Errors (discordant calls)
= 366 Remaining Errors
= 0.13 errors per called MB

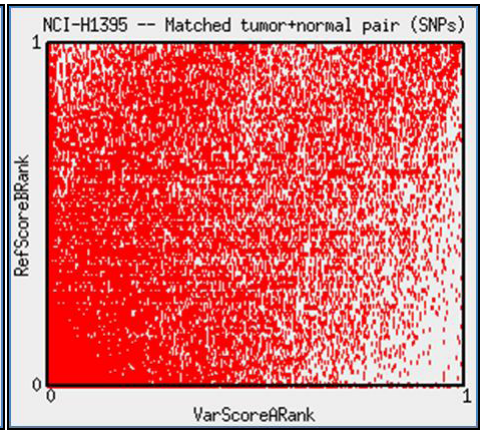
10x Improvement in Sensitivity/Specificity Trade-Off. Quantitative comparisons are more powerful than single sample analysis

Krishna Pant: NA12878 Replicate Data Analyzed Using Development Pipeline. 45

Distribution of somatic variant calls

- Concordant Calls (likely True+s) and discordant calls (False somatic events) have a different distributions of strength of evidence.





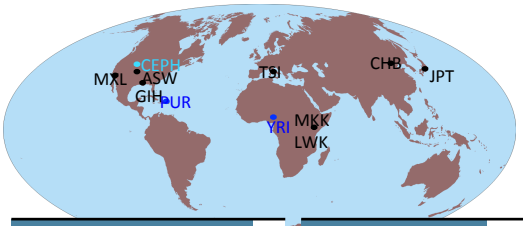
46

Case Studies of SNP Validations in Tumors

Genentech 1	• Moderately highly mutated NSC lung tumor with ~50K True+ • Moderately aneuploid, matched normal is margin • 90% Validation rate using (old version of) CG Somatic Score *Lee et al., Nature 2010*
Genentech 2	• Lung Cancers from non-smokers – 10x lower True+ • "Similar" validation rates using newer Somatic Score *Zhang AGBT 2011*
Customer S	• Solid Tumor with very low True+ rate, <<1000 genome-wide • Minimal CNV/SVs seen • 17 for 25 Somatic SNP validation rate in spite of low True+ *Pers. Comm*
Customer T	• Blood cancer with tightly matched phenotypes • Identical activating mutation found in 90% of tumors • Negative had mutation in 10% of reads (highly mixed sample) *Manuscript in preparation*
Erasmus MC	• Blood cancer using post-therapy remission samples as normals • 91% validation rate on SNPs and small dels *J Meijerink; UGM 2011*

A Reference Panel of 69 Whole Human Genome Sequences

Population Diversity of 69 Reference Genomes

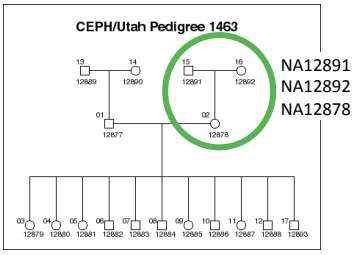


*MXL - Mexican ancestry in Los Angeles, California	5	CEU - Utah residents with Northern and Western European ancestry from the CEPH collection	5	*ASW - African ancestry in Southwest USA	5	CHB - Han Chinese in Beijing, China	4
*Puerto Rican Trio	3	CEPH – 17 member 3 generation	17	LWK - Luhya in Webuye, Kenya	4	JPT - Japanese in Tokyo, Japan	4
		TSI - Tuscans in Italy	4	MKK - Maasai in Kinyawa, Kenya	4	*GIH – Gujarati Indian ancestry in Los Angeles	4
				YRI - Yoruba in Ibadan, Nigeria	7		
				Yoruban Trio	3		

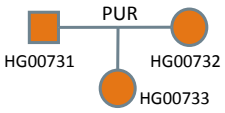
*Significantly Admixed Populations

Complete Genomics Deeply Sequenced Public Genome Release

17 Member 3 Generation Nuclear Family
- CEPH/Utah Pedigree 1463
- 4 Grandparents, 2 Parents, and 11 Children
- Includes deeply sequenced trio from the 1000 Genomes Project

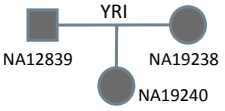


Yoruban Trio (mother, father, child)
- Also deeply sequenced by 1000 Genomes



Puerto Rican Trio (mother, father, child)
- DNA is from a relatively young cell-line

46 Ethnically diverse unrelated samples
- All from HapMap and/or 1000 Genomes



completegenomics.com bionumbus.org dnanexus.com

CGA Tools™ Open Source Software cgatools.sourceforge.net

Function	Description
map2sam	Convert initital mappings to SAM (and thus BAM)
evidence2sam	Convert de novo assemblies to SAM (and thus BAM)
generatemastervar	Create materVar file from CG genome (easy to use)
snpdiff	Compare SNP genotypes to CG genome*
calldiff	Compare two CG genomes, optionally compute Somatic Score*
testvariants	Compare multiple CG genomes*
listvariants	Prepare genomes for testvariants*
join	Add additional annotations (columns) to CG files
fasta2crr	Create CG format reference database from FASTA files
crr2fasta	converts CRR sequence files to FASTA file format
decodecrr	retrieves the sequence for a given range of a chromosome
listcrr	Lists chromosomes, contigs, ambiguous regions
help	Display on-line help for CGA Tools command

* Take into account complex variations, partial information, etc.

High Concordance Between CG Sequences and Public Data for 69 Public Genomes

Description		Median%	Range
Genome call-rate(reference or variation, not no-call)		96.81	95.6 - 97.4
Exome call-rate		95.92	93.9 - 96.9
Hapmap 1/2 Infinium HQ Subset	called by CGI	99.32	97.13 - 99.51
	called *concordantly* by CGI	**99.94**	99.88 - 99.96
Hapmap 3	called by CGI	99.45	97.77 - 99.66
	called *concordantly* by CGI	99.73	99.37 - 99.76
1000 Genomes Low Pass SNP loci*	called by CGI	98.73	96.94 - 99.46
	called *concordantly* by CGI	99.83	91.46 - 99.18
1000 Genomes High Depth Trios SNP loci**	called by CGI	99.71	97.59 - 99.78
	called *concordantly* by CGI	**99.59**	99.41 - 99.70

* Published FDR: 3.3-4.3%
** Published FDR: 1.5-2.4%

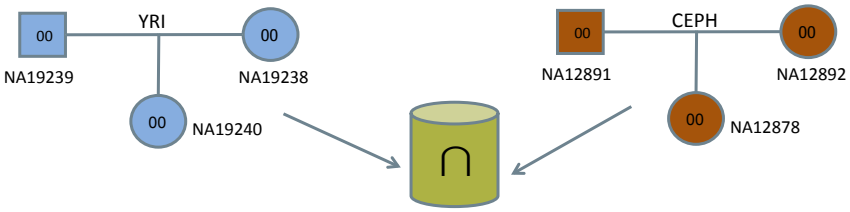
Complete Genomics published FDR: 0.2-0.6%
(older data, more recent data are better still)

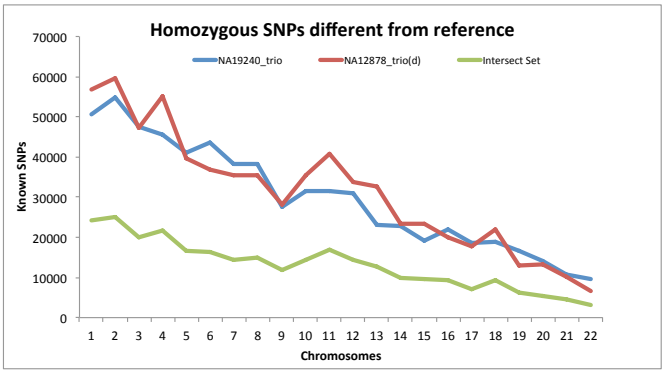
Comparisons using Open Source CGA Tools Software v1.3: cgatools.sourceforge.net

Discordances of CG vs. 1000 Genomes Project: Validation by Sanger sequencing

Discordant Novel SNP Loci 1KG= 1000 Genomes Project	Successfully Sequenced Loci	CGI Validation Rate	1000 Genomes Validation Rate
CG: Heterozygous SNP/Reference 1KG: no SNP	46 (of 79,381)	95.6%	-
CG: Homozygous SNP 1KG: no SNP	45 (of 5,638)	93.3%	-
CG: no-call 1KG: Heterozygous SNP	36 (of 2,962)	-	94.4%
CG: Homozygous Reference (e.g. no snp) 1KG: Heterozygous SNP	88 (of 403)	74.2%	22.5%

Rare reference alleles or reference errors?

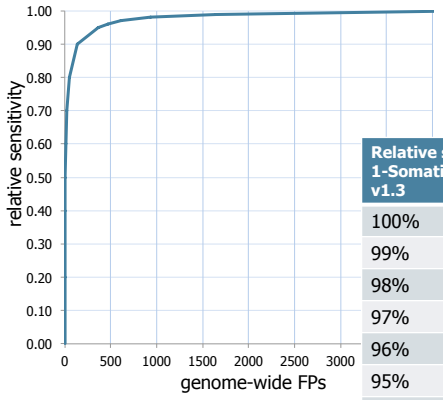




Summary

- High-depth high-accuracy whole human genome sequencing is becoming practical on a large scale
- Specialized methods for variant calling can elucidate the complexities of germ-line and somatic variation in human genomes
- Reference data sets are available to give one a good handle
  - E.g. 69 Genomes at completegenomics.com
- Tools for downstream analysis and annotation of genomes are fast becoming ***the*** a critical area for the field
  - E.g. CGA Tools on cgatools.sourceforge.net
  - Bioconductor

Performance of Somatic Score v1.3: Controlling Sensitivity/Specificity Trade-off



Relative sensitivity 1-Somatic Score v1.3	Genome-wide False+ somatic SNPs	Exome-wide False + disruptive somatic SNPs
100%	4258	24
99%	1649	8
98%	937	5
97%	608	2
96%	464	2
95%	366	2
94%	273	1

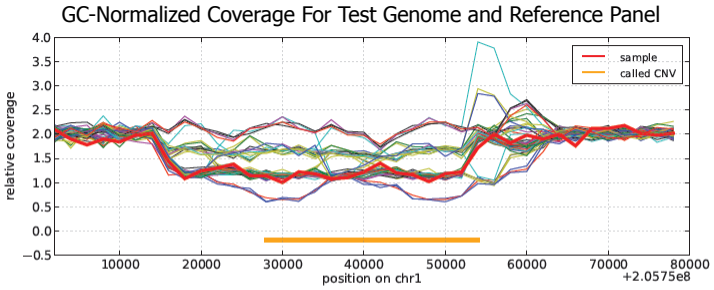
Complete genomics

Thank You

Complete genomics

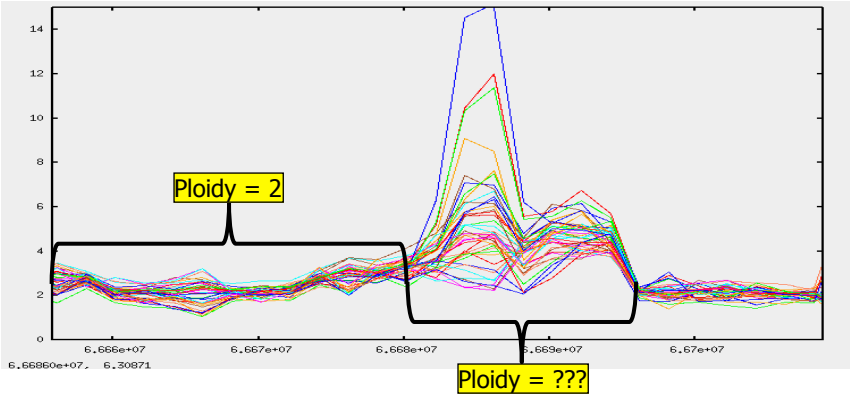
Extra Slides

The Challenge: Hypervariable CNV Sites in our Reference Panel



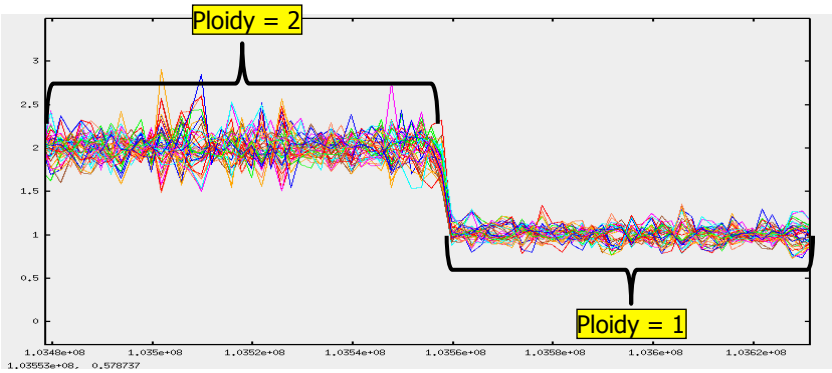
- Region is a known SegDup (~100% overlap)
- In this case, clustering of baseline genomes will likely result in region being called diploid or CNV in most genomes of interest by CG's algorithm

A More Difficult Hypervariable Region In The Reference Genome



- Some regions of the reference show highly variable coverage which does not cleanly cluster into discrete ploidies
- May be high copy repeat regions, highly polymorphic CNV, or sequencing or mapping artifacts
- May be labeled "hypervariable" and no ploidy assigned by CG's HMM

Another Challenge: Invariant Regions In the Reference Genome

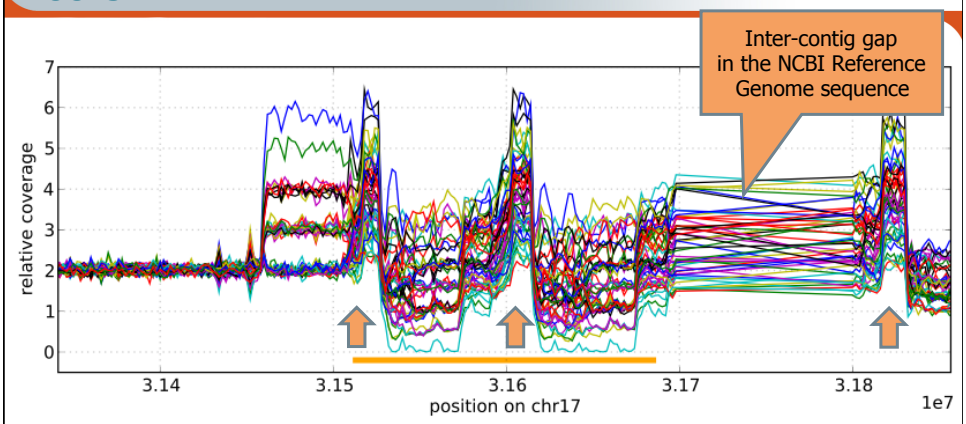


- Coverage consistently indicates the same CNV (loss) across many samples
- Likely a duplication in the reference genome not in any of the reference samples
- Labeled "invariant" and no ploidy assigned by the CG HMM

CCL3

- The copy number of CCL3 varies among individuals. Most individuals have 1-6 copies in their diploid genomes, although rare individuals have zero or more than six copies.
- The human genome reference assembly contains two full copies of the gene (CCL3L1 and CCL3L3) and an additional partial duplication, which is thought to result in a pseudogene, CCL3L2.
- Many platforms seem to struggle with this region:
  - The CCL3 sequence aligns at 2 sites on reference genome chr17 and one on chr17_random (unincorporated sequence from the Human Genome Project). The human genome reference has a large inter-contig sequence gap of N's.
  - HapMap indicated no CNVs flanking this gene. Likely false negative.
  - Conrad et al. found a Copy Number loss in 2/19 CEU individuals using (only) the Nimblegen Array. Also likely false negative?

CCL3L1



Shows signal both GC normalized and also normalized against the reference panel.

Normalizing against the reference panel assumes that our ploidy estimate of each reference sample. This may explain the shifts highlighted.

How to Attack Complex CNV Sites in Reference Panel

- Complete Genomics provides Base by Base GC-Normalized coverage data in the coverageRefScore files
  - E.g. Use data prior to normalization against reference panel
  - Select appropriate smoothing window (1k-10k) and method
  - Do direct comparisons vs. carefully selected "normal" control(s) in region(s) of interest
    - Either Scale GC coverage data to ploidy 2
    - Or Select scaling factors which minimize R^2
  - Various R CNV packages have been used with CG data successfully post smoothing

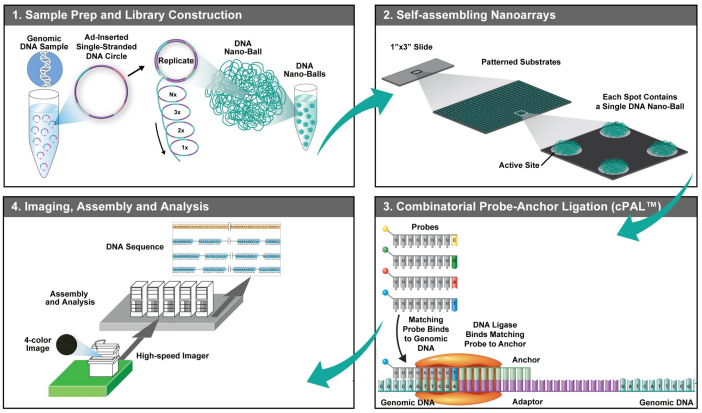
Drmanac et. al., Science, 2010

Science 1 January 2010:
Vol. 327. no. 5961, pp. 78 – 81
DOI: 10.1126/science.1181498

Science

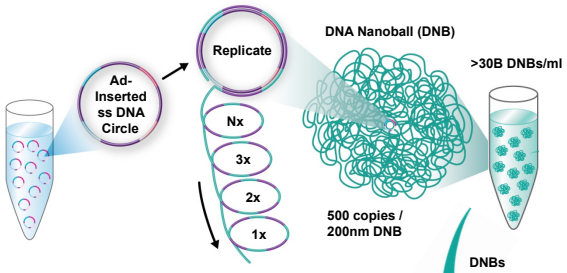
REPORTS

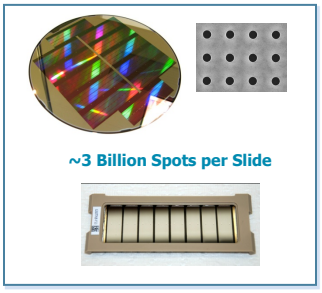
Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays

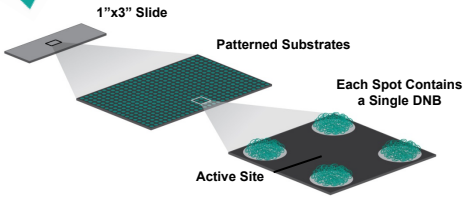


DNA Nanoball (DNB) Formation and Array Preparation

Reaping the benefits of single molecule processing without the cost of single molecule detection

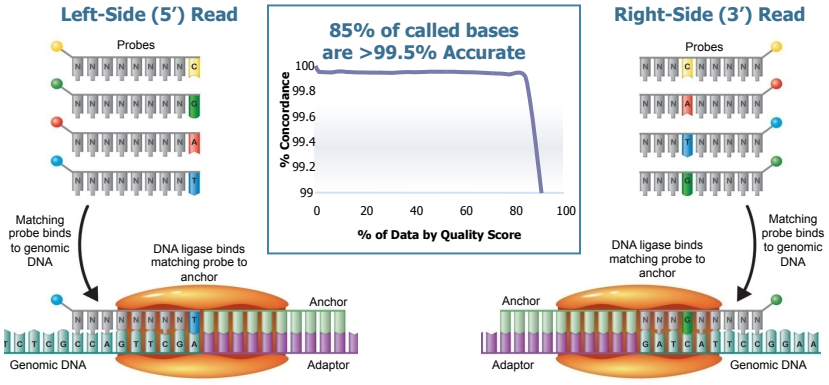






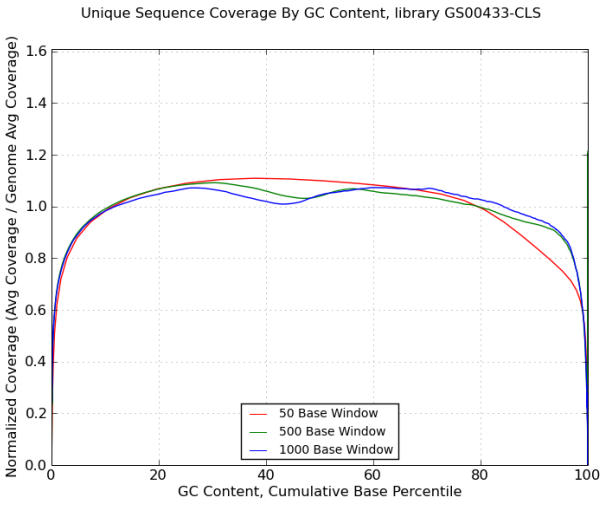
- Low cost to make (in solution)
- Low cost to assay (small volume)
- Low cost to image (2 pixels/spot)

Combinatorial Probe-Anchor Ligation (cPAL): Unchained Base Reading



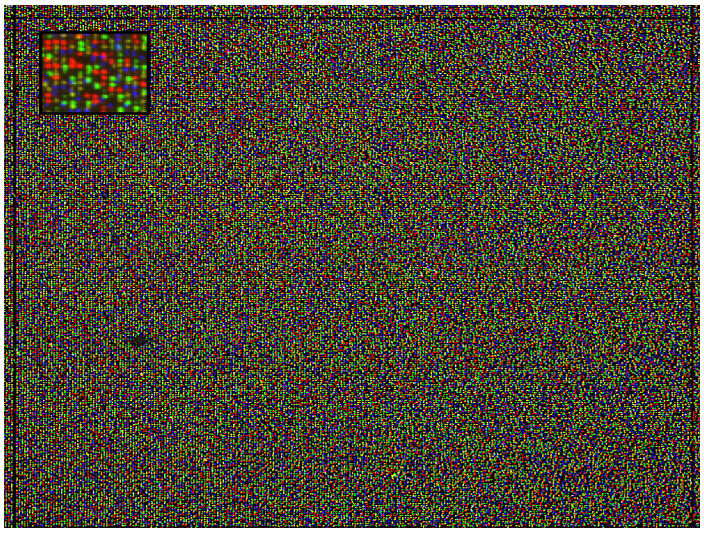
- Accurate unchained base reads from multiple start points (extendable read length): each sequencing reaction is independent of previous reaction
- Reduced systematic errors by sequencing from 2 directions; 5 different probe reagent sets
- Low concentration of simple labeled probes (low cost)

NA19240: Unique Coverage of Clones and Sequence reads by GC%

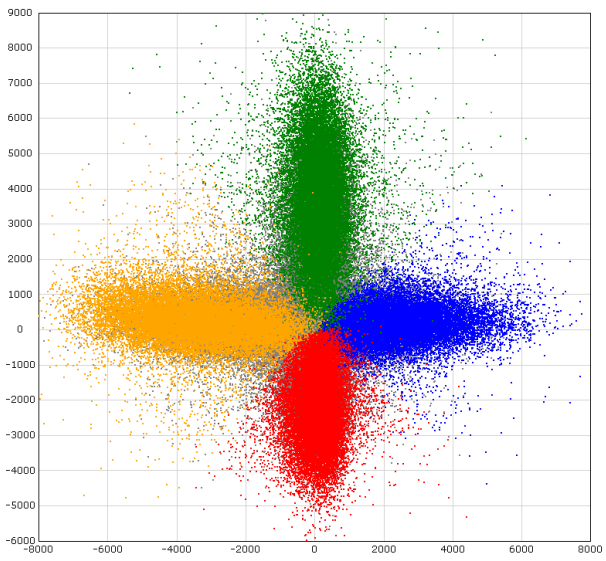


NA19240 data from Drmanac et al. Science 2009

Four-color Image of a DNB Array Field



Successful Base Calling Using Only 2 pixels per DNB



Extensive QC/QA and Process Controls

- **Pervasive LIMS System throughout the process**
- **Standardized Testing of:**
  - **Samples upon receipt**
  - **Instruments and Software Releases**
  - **Reagent lots (including Slide lots)**
  - **Individual library construction intermediates**
  - **DNA Nano Balls**
- **Automated QC of Each Data Set**
  - **Extensive set of computed metrics**
  - **Track historical performance on these metrics in our LIMS**
- **Details and Data can be provided in the future**

A Few Helpful Notes on NGS Metrics

- **CG's analysis process is different than many others**
  - Unfortunately, one can't look at certain intermediate outputs from the CG pipeline output as if they were from some different pipeline or platform. For example: pre vs. post *de novo* assembly coverage changes significantly.
  - Coverage metrics rarely properly consider the repetitive structure of approximately half of the human genome. Example: The most commonly used algorithms for other platforms **randomly** place reads which map to high copy repeats. CG does not do that adding to perceived gaps to our coverage.
- **Read level metrics can be misleading indicators of performance**
  - For example: Read error rates only directly indicate consensus error rates **if and only if** those read-level errors are truly random
  - Edge effects dominate many counting statistics allowing one to be easily misled. Example: Gaps between CG sub-reads in the very small fraction of the genome which gets very low mapped read coverage are a large fraction of the count of gaps in total.
- **One is usually best off looking at net performance**

Complete Genomics: What's New

- **Greatly expanded sample and library QC, reduced library biases**
  - More even coverage and higher call rates over the genome (avg ~96%)
- **Improved Biochemistry and Small Variant Calling Algorithms**
  - Improved accuracy: ~3 errors per MB
- **New Calling Algorithms and Annotations**
  - masterVar file (much easier to parse)
  - CNVs (quantitative, using read-depth) and SVs (paired-end analysis)
  - Mobile Element Insertions (paired end analysis)
  - Re-written protein and dbSNP annotation pipeline: RefSeq 37.2 dbSNP 132
  - microRNAs, PFAM, COSMIC, GRC build 37, dbSNP 132
- **CGA Tools 1.4 Open Source Software**
  - Improved AUC in somatic score 1.3-1.4
  - Pairwise and multi-genome comparisons of small variants
  - junctionDiff and junctions2events
- **Public Release of 69 Genomes**

A Critical Point on Measuring Accuracy (on any sequencing platform)

- Concordance with SNP genotypes at known sites of common variation is not the same as overall sequence accuracy

1. SNP arrays (and databases of known SNPs) are highly biased toward "easy" regions of the genome
2. False Positives in Homozygous-Reference Regions Are Not Measured, But Must Be

- Overall Error Rate and False Discovery Rate Are Not the Same
  - Imagine a 1:100,000 bp genome-wide error rate
  - Imagine that 50% errors are false positive and 50% are false negative
  - That would be one false positive every 200,000 bp
  - There is a true positive SNP very roughly every 1,000 bp
  - One thus expects an FDR of roughly 0.5% (1/201)