Knowing Your NGS Upstream: Alignment and Variants

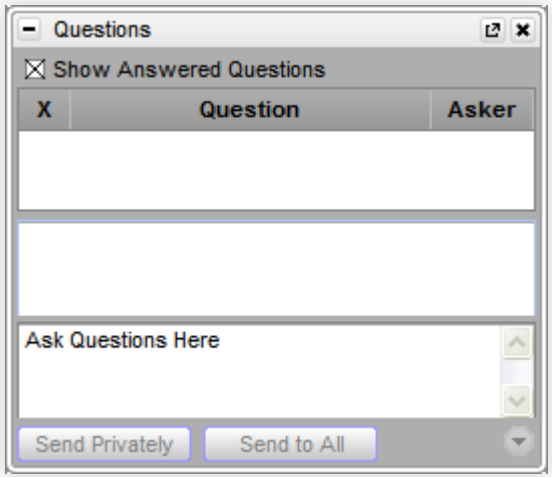
March 27, 2013

Gabe Rudy, Vice President of Product Development

GOLDEN HELIX
Accelerating the Quest for Significance™

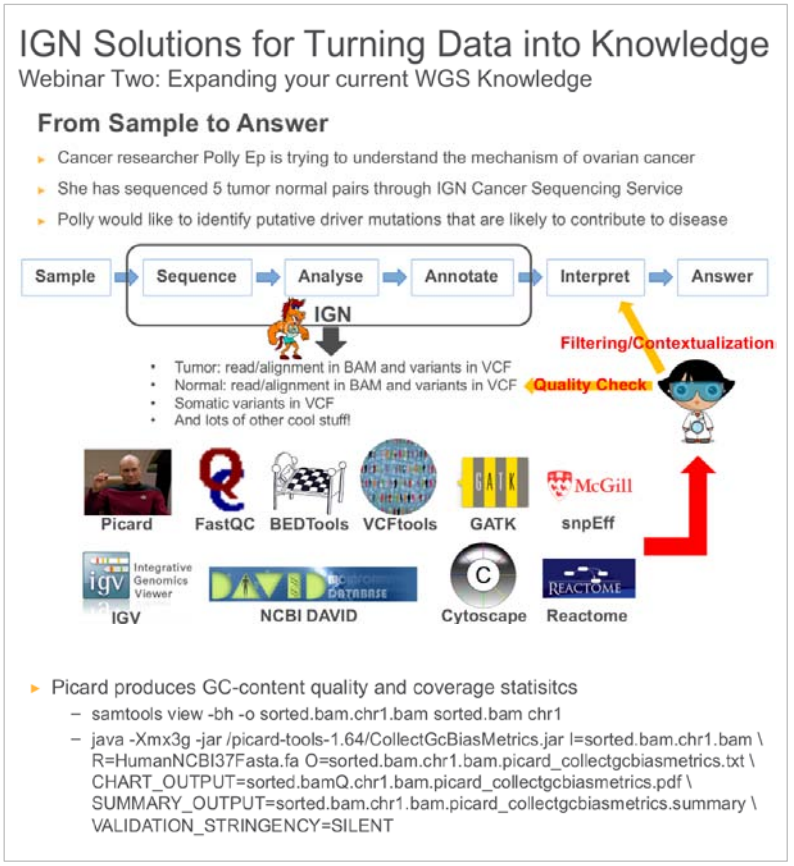
Questions during the presentation

Use the Questions pane in your GoToWebinar window



Goals

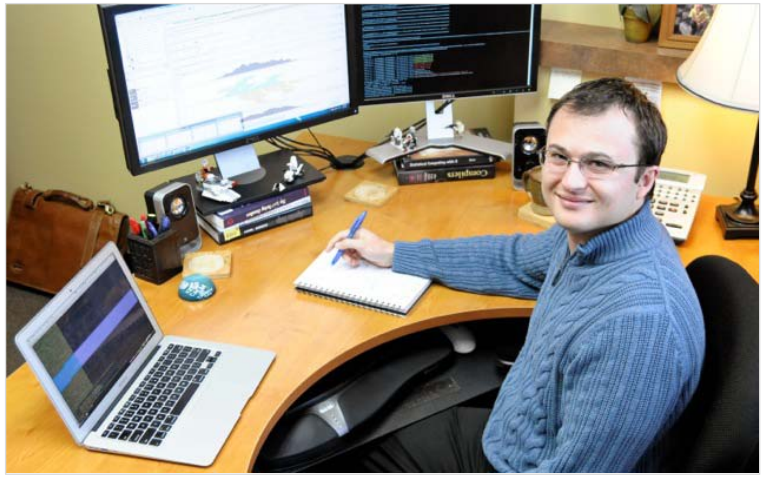
- **What I Assume About You**
  - Some experience with NGS technology
  - Not a command line bioinformatician by day; not afraid of technical terms
- **What You Will Learn**
  - A healthy skepticism when looking at NGS data
  - What to expect/not expect from core labs or upstream sequencing service providers
  - Reading pile-ups in a genome browser and spotting high quality vs sketchy variants
- **What You Won't Learn**
  - Interpreting biological significance of variants
  - One true way to do secondary analysis



I won't do this… hopefully!

My Background

- **Golden Helix**
  - Founded in 1998
  - Genetic association software
  - Analytic services
  - Hundreds of users worldwide
  - Over 700 customer citations in scientific journals
- **Products I Build with My Team**
  - **SNP & Variation Suite (SVS)**
    - SNP, CNV, NGS tertiary analysis
    - Import and deal with all flavors of upstream data
  - **GenomeBrowse**
    - Visualization of everything with genomic coordinates. All standardized file formats.
  - **RNA-Seq Pipeline**
    - Expression profiling bioinformatics



Agenda

1. Background and Definitions
2. Why You Should Care About Your Upstream?
3. A Drink from the Bioinformatics Firehose
4. Service Provider Deliverables: CEPH Trio Example
5. Applications That Require Special Upstream Analysis

Agenda

1. **Background and Definitions**
2. Why You Should Care About Your Upstream?
3. A Drink from the Bioinformatics Firehose
4. Service Provider Deliverables: CEPH Trio Example
5. Applications That Require Special Upstream Analysis

NGS Analysis

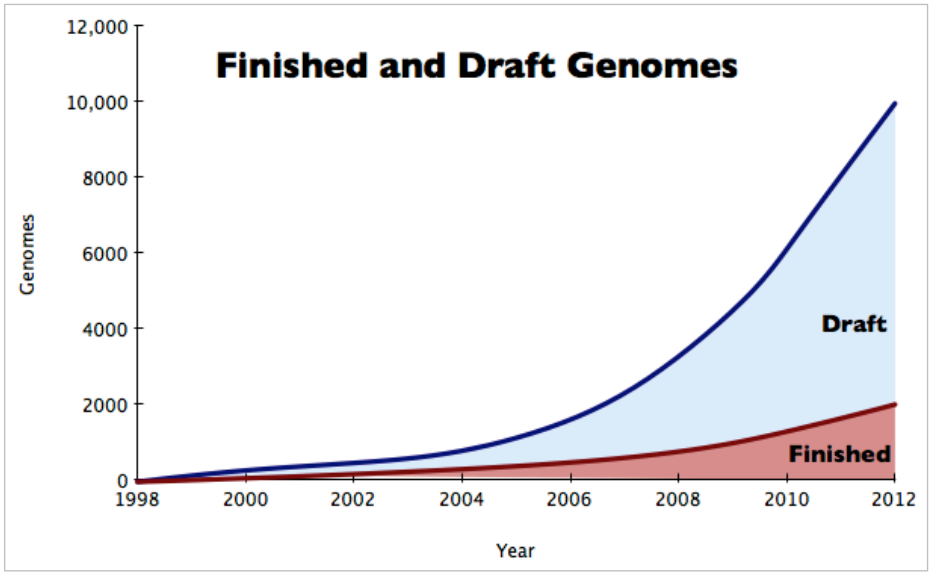
Primary Analysis	▪ Analysis of hardware generated data, on-machine real-time stats. ▪ Production of sequence reads and quality scores
Secondary Analysis	▪ QA and clipping/filtering reads ▪ Alignment/Assembly of reads ▪ Recalibrating, de-duplication, variant calling on aligned reads
Tertiary Analysis **“Sense Making”**	▪ QA and filtering of variant calls ▪ Annotation and filtering of variants ▪ Multi-sample integration ▪ Visualization of variants in genomic context ▪ Experiment-specific inheritance/population analysis

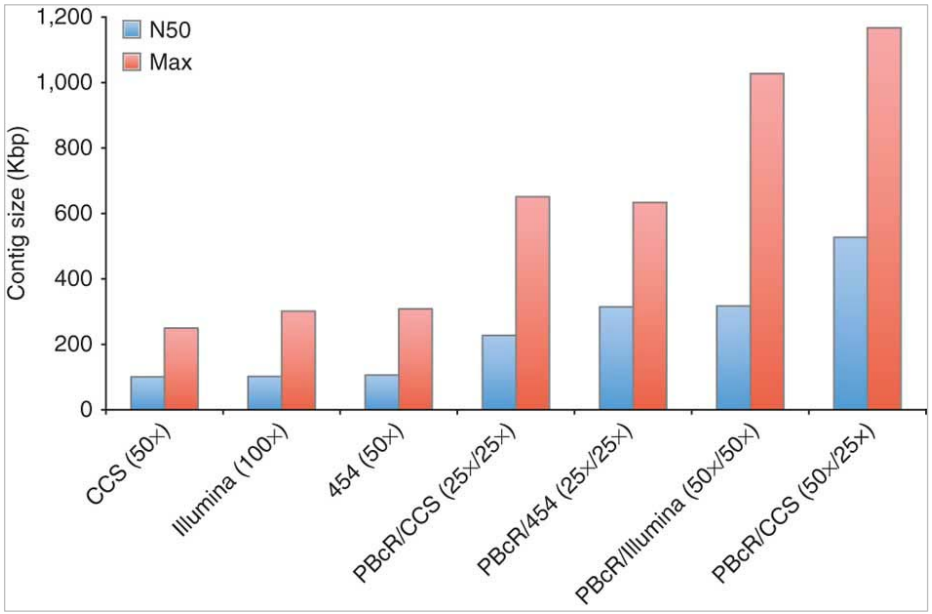
Primary Analysis: I'd like some AGCT's please

- **Standardized on producing FASTQ**
  - AGCT or N
  - Quality scores
  - Pair of files for paired end
- **Happens on machine for desktop sequencers**
  - Ion Torrent processing microwell detectors
  - MiSeq doing optic processing of flowcell
  - PacBio processing optics of ZMW
- **HiSeq 2000/2500**
  - Requires off-machine base-calling
  - Can "call bases" with Illumina software on raw data collected tile by tile

Assembly vs Alignment

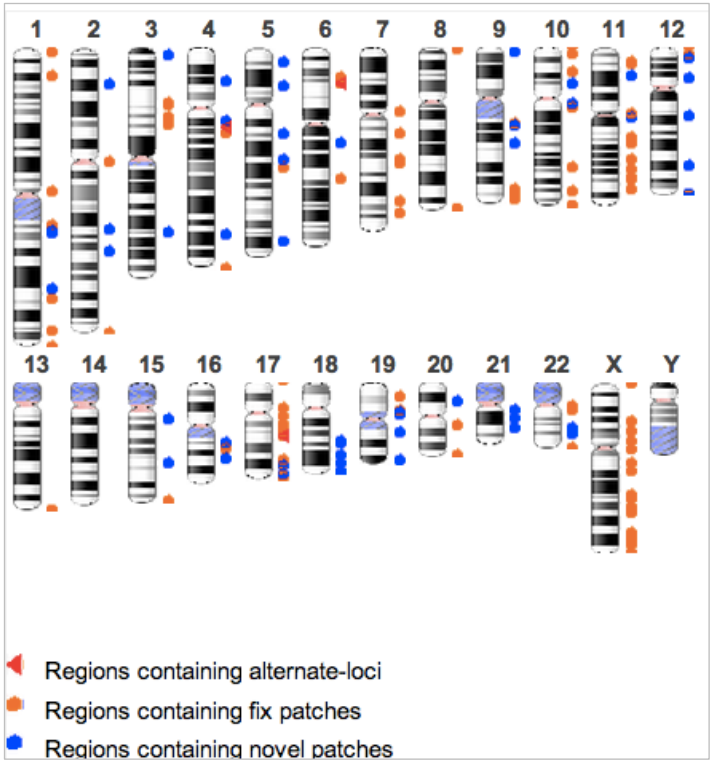
- **De Novo Genome Assembly**
  - Very difficult for large genomes to get to “finished” genome quality (traditionally done with Sanger).
  - Short reads will get you to contigs sizes of ~10-100Kb range.
  - Need long reads (PacBio) or restriction maps optical mapping (OpGen) to make chromosomal sized contigs
- **Alignment**
  - Aligning to finished (or draft) genomes that is considered “reference”
  - Allows for some differences, but not too many between your reads and the reference





The Human Reference Sequence

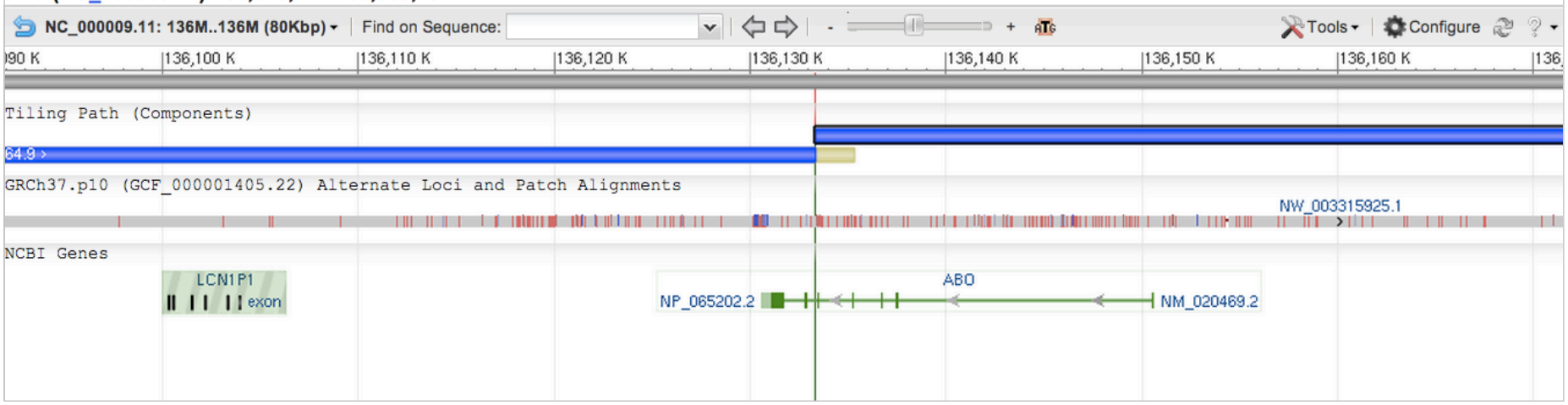
- **Genome Reference Consortium (GRCh37)**
  - Feb 2009, previous was NCBI36 March 2006
  - 9 alt loci and 187 patches (11 patch releases)
- **Supercontigs: Large unplaced contigs**
  - Some localized to chr level and some unknown
- **Does not include a Mitochondrial reference**
  - UCSC hg19 includes older NCBI 36 MT
  - 1000 genomes project using revised Cambridge Reference Sequence (rCRS)
  - Provide “g1k” reference: includes rCRS, Human herpesvirus 4 type 1, supercontigs and “decoy” sequence
- **v38 genome coming this summer:**
  - Incorporate all patches into the reference
  - Some allele fixes to have reference match major



Example Patch

GenBank ID	RefSeq ID	Sequence Type	Length	Alignment mismatch	Unique sequence	Show in Viewer
GL339450.1	NW_003315925.1	Fix	330,164	500	10,751	○

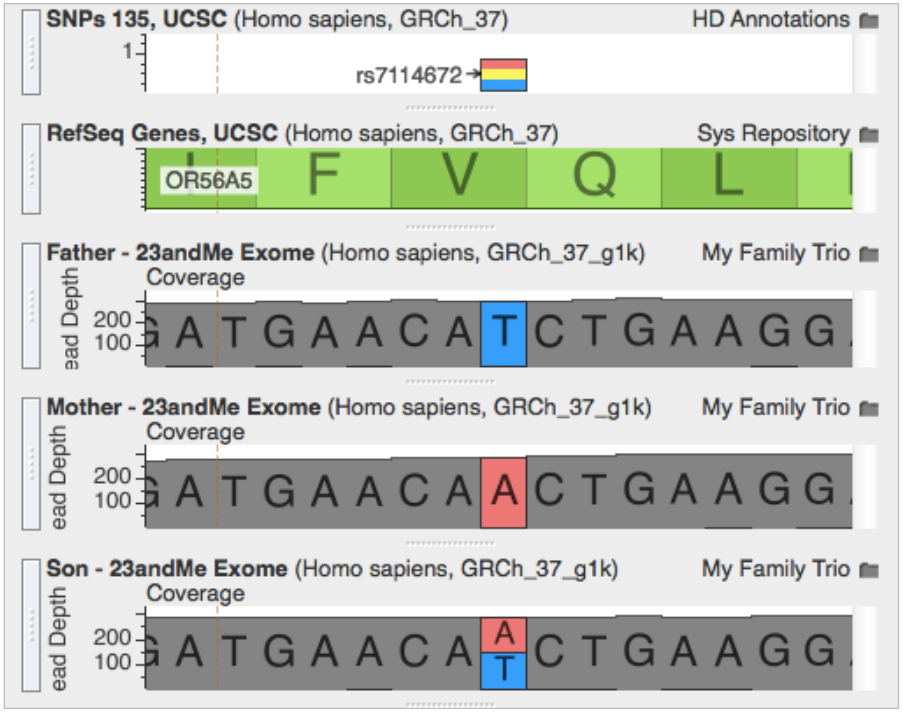
◉ Show region

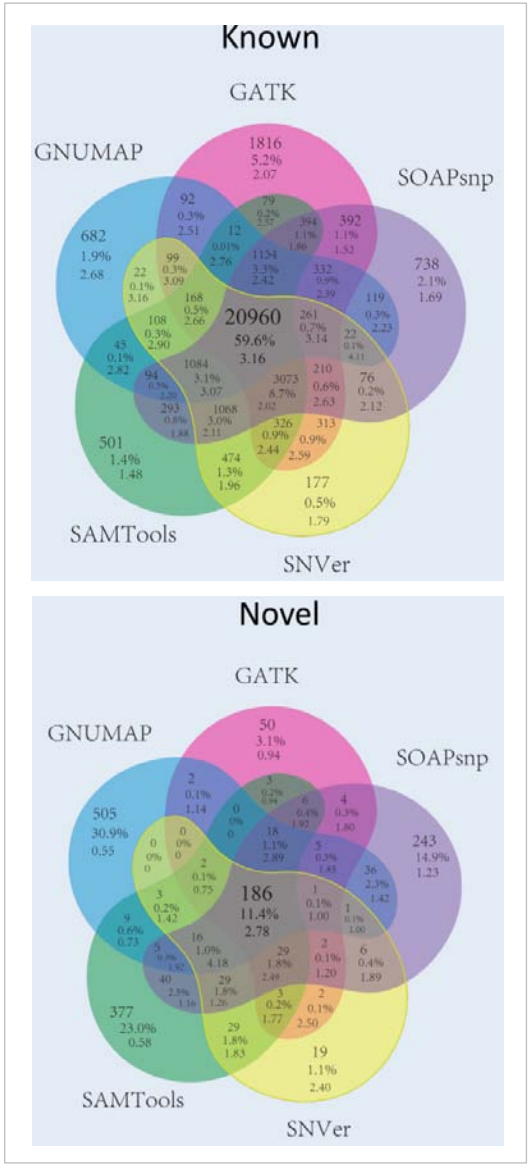


- **The tiling path in GRCh37 switched in the middle of ABO gene** resulting in a reference protein not present in humans.
- **Patch adjusts tile path and fixes the problem.**
- **All patches will be incorporated into GRCh38, due this summer.** Until then, all alignment is done against unpatched reference.

Single Nucleotide Variants (i.e. SNVs or SNPs)

- Single base substitution from reference
- Note that “reference” is not always the “major” allele
- “Multi-allelic” sites have more than 2 cataloged alleles



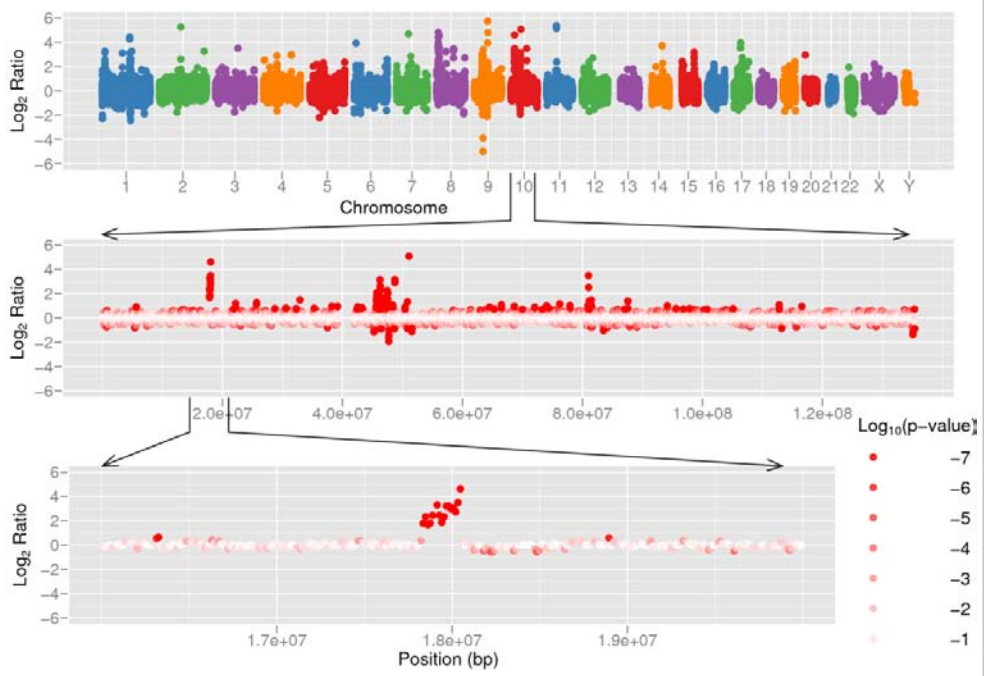


Small Insertions/Deletions

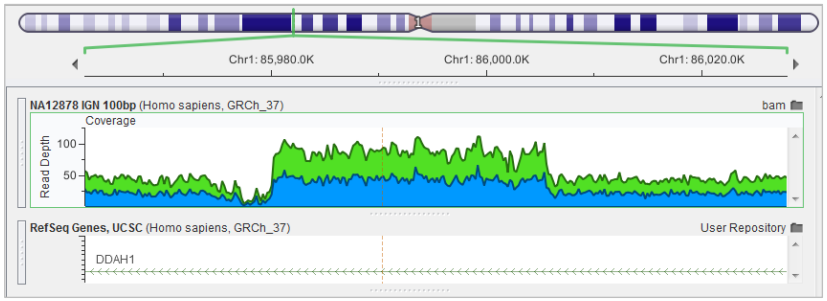
- **Generally defined as being < 150bp (often much shorter)**
- **Frameshift insertions/deletions important “loss of function” class of variants**
  - Although InDels divisible by three are “in-frame” when in coding region
- **Hard to call consistently. Poor concordance between algorithms.**
- **Where to call an InDel in a homopolymer?**
  - GTTTAC
  - GTTTTAC
  - 01234567
  - How do you describe the insertion? Ins of T at 5? Or ins of T at 1?
    - CGI in their v1 pipeline preferred calling insertion at end, others at beginning, now always at beginning
- **MNP – Can also be called differently**

Copy Number Variants

- **Required WGS**
  - CNVs > 10kb pretty accurate.
  - 1kb to 10kb problematic.
- **Detecting Deletions**
  - Can see coverage drop to near zero
  - Harder to pinpoint breakpoint
  - Possible false positives in low-mappability regions
- **Amplifications**
  - Can see coverage jump
  - False positives due sample prep or sequence artifacts
- **Need "baseline," look at Log Ratio**
  - Somatic detection uses normal tissues
  - Can have control population



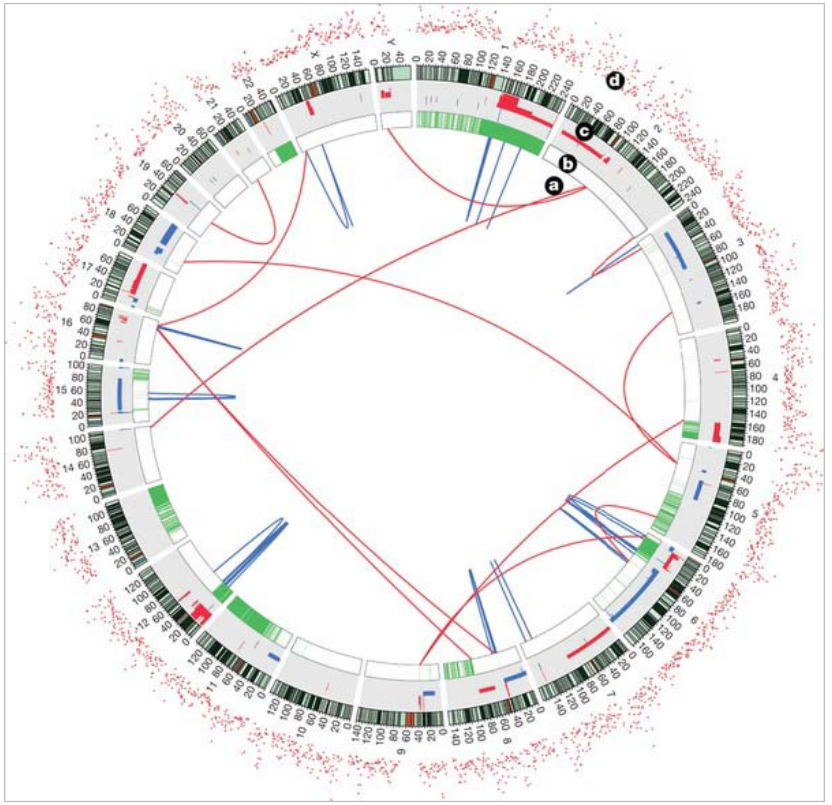
Venter vs Watson WGS CNV-seq



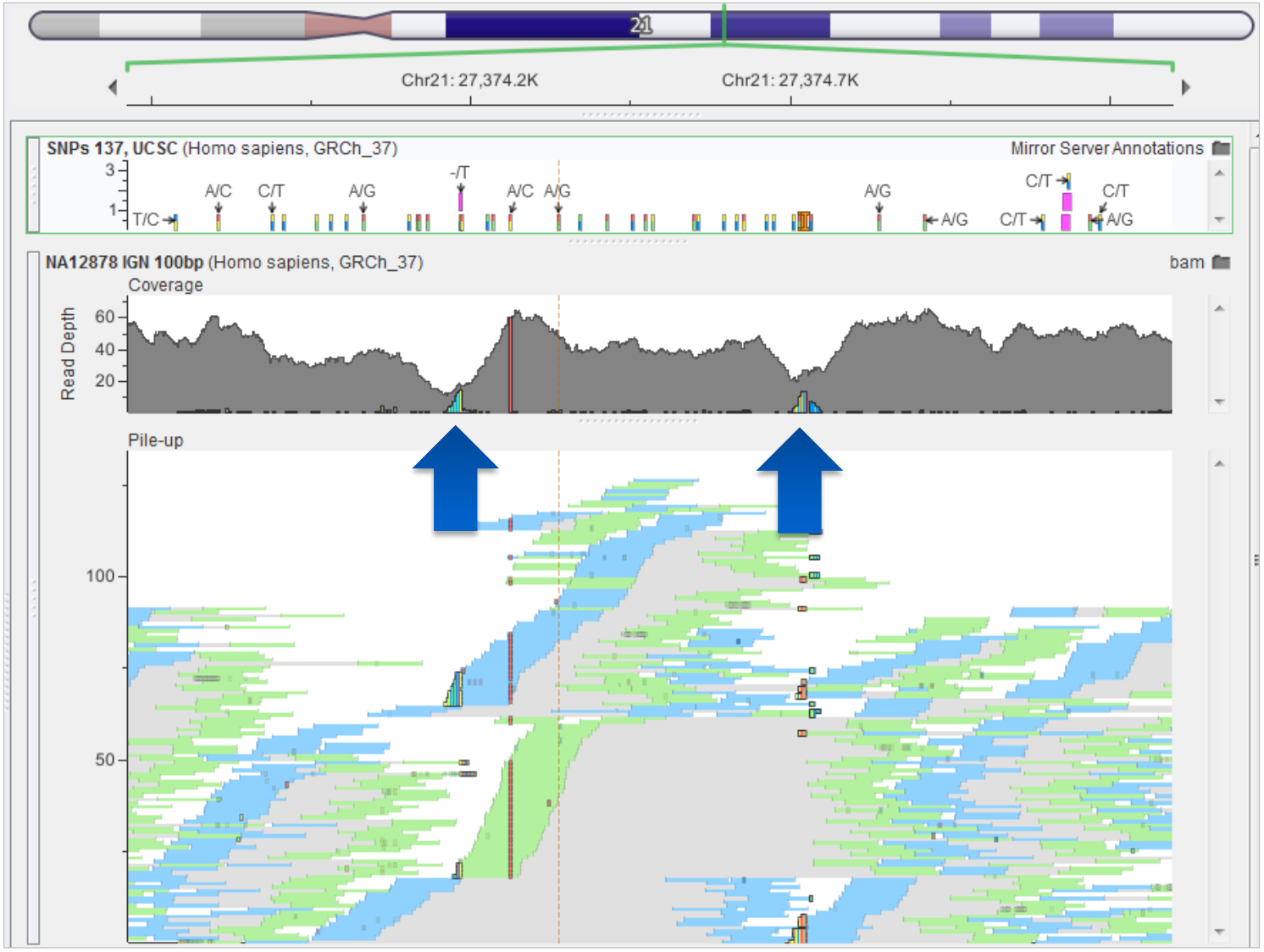
64kbp gain in DDAH1 Gene of NA12878

Structural Variants

- **Looking for:**
  - Balanced rearrangements
  - Inversions
  - Translocations
  - Complex
- **Signals to detect SV:**
  - Paired-end mappings too big (deletion)
  - Depth of coverage
  - Split-read mapping
- **Translocations can result in “fusion” genes.**
  - For example BCR-ABL fusion gene central in pathogenesis certain leukemias.



Example 1kb Inversion (intron of APP)



Tertiary Analysis – “Sense Making”

- **Detecting Known Clinically Relevant Variants**
  - Use targeted gene panels. Amplicons or custom capture.
  - Look for carrier status or present of pathogenic or PGx variants
- **Rare, Functional Variant Search and Interpretation**
  - Rare Mendelian Diseases
  - Clinical Diagnostic: Ending Diagnostic Odyssey
  - Looking for rare variants of functional consequence to a known phenotype
  - Exome sequencing common, but whole genome has proponents
  - Trios often used for looking at inheritance of putative variants (compound hets)
- **Population Studies**
  - Like NHLBI or others studying complex disease
  - Often looking at “variant burden” over genes between cases/controls
- **Driver Somatic Variant Identification**
  - Looking for variants in tumor samples but not matched normal
  - Not just SNPs and InDels, but CNVs and SVs

Not just DNA… but still DNA sequencing

- **RNA-Seq**
  - Align to “transcriptome”, but often do analysis with reference genome coordinates and reads “gapped” over introns they span in their spliced form
  - Using read counts to approximate relative abundance of RNA in sample
  - Compare relative abundance between groups
  - Discover new transcribed genes or alternative splicing
- **ChIP-Seq**
  - Measure sites and intensities of various proteins binding to DNA
  - ENCODE project used to catalog TFBS and other functional elements
- **Methyl-Seq**
  - Get sequences only with epigenetic methylation mark
  - Run peak identification and intensity to look at relative levels of methylation

Agenda

1. Background and Definitions
2. **Why You Should Care About Your Upstream?**
3. A Drink from the Bioinformatics Firehose
4. Service Provider Deliverables: CEPH Trio Example
5. Applications That Require Special Upstream Analysis

The Promise

- **Both in research and clinical care**, NGS is powering discoveries making impactful diagnoses
- **Desktop sequencers and gene panels much more economical** than gene-by-gene hunts
- **Exomes have lead to many rare disease diagnoses** and affordably assay rare functional variants
- **Whole genomes have led to clinical success** stories and promise to be instrumental to our understanding of complex disease genetics
- **Barrier to entry is lower than ever**

Things That Can Confound Your Experiment

Library preparation errors	Sequencing errors	Analysis errors
▪ PCR amplification point mutations (e.g. TruSeq protocol, amplicons) ▪ Emultion PCR amplification point mutations (454, Ion Torrent and SOLiD) ▪ Bridge amplification errors (Illumina) ▪ Chimera generation (particularly during amplicon protocols) ▪ Sample contamination ▪ Amplification errors associated with high or low GC content ▪ PCR duplicates	▪ Base miscalls due to low signal ▪ InDel errors (particular PacBio) ▪ Short homopolymer associated InDels (Ion Torrent PGM) ▪ Post-homopolymeric tract SNPs (Illumina) and/or read-through problems ▪ Associated with inverted repeats (Illumina) ▪ Specific motifs particularly with older Illumina chemistry	▪ Calling variants without sufficient reads mapping ▪ Bad mapping (incorrectly placed read) ▪ Correctly placed read but InDels misaligned ▪ Multi-mapping to repeat/paralogous regions ▪ Sequence contamination e.g. adaptors ▪ Error in reference sequence ▪ Alignment to ends of contigs in draft assemblies ▪ Incorrect trimming of reads, aligning adaptors ▪ Inclusion of PCR duplicates

Nick Loman: Sequencing data: I want the truth! (You can't handle the truth!)
Qual et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 Jul

Your Choice of Technologies, Sometimes…

Platform	Illumina MiSeq	Ion Torrent PGM	Ion Torrent Proton	PacBio RS	Illumina HiSeq 2000
Instrument Cost*	$125 K	$50 K	$150K	$695 K	$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316, 1Gb on 318	10Gb on PI, 30GB on PII (Mid 2013)	100 Mb	600Gb
Sequencing cost per Gb*	$502	$500 (318 chip)	$70 (PI chip)	$2000	$41
Run Time	27 hours***	2 hours	3 hours	2 hours	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	Claimed >Q30	<Q10	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	Probably ~1%	12.86 %	0.26 %
Read length	up to 150 bases	~200 bp	100bp (200bp PII)	Average 1500 bases	up to 150 bases
Paired reads	Yes	Yes	Yes	No	Yes
Insert size	up to 700 bases	up to 250 bases	up to 250 bases	up to 10 kb	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	100-1000 ng	~1 µg	50-1000 ng
Applications	Targeted	Targeted	Exomes, RNA-Seq	Assembly, Validation	Exomes, Genomes, RNA

Qual et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012 Jul 29

[Show SNPs/Indels GenomeBrowse]

Agenda

1. Background and Definitions
2. Why You Should Care About Your Upstream?
3. **A Drink from the Bioinformatics Firehose**
4. Service Provider Deliverables: CEPH Trio Example
5. Applications That Require Special Upstream Analysis

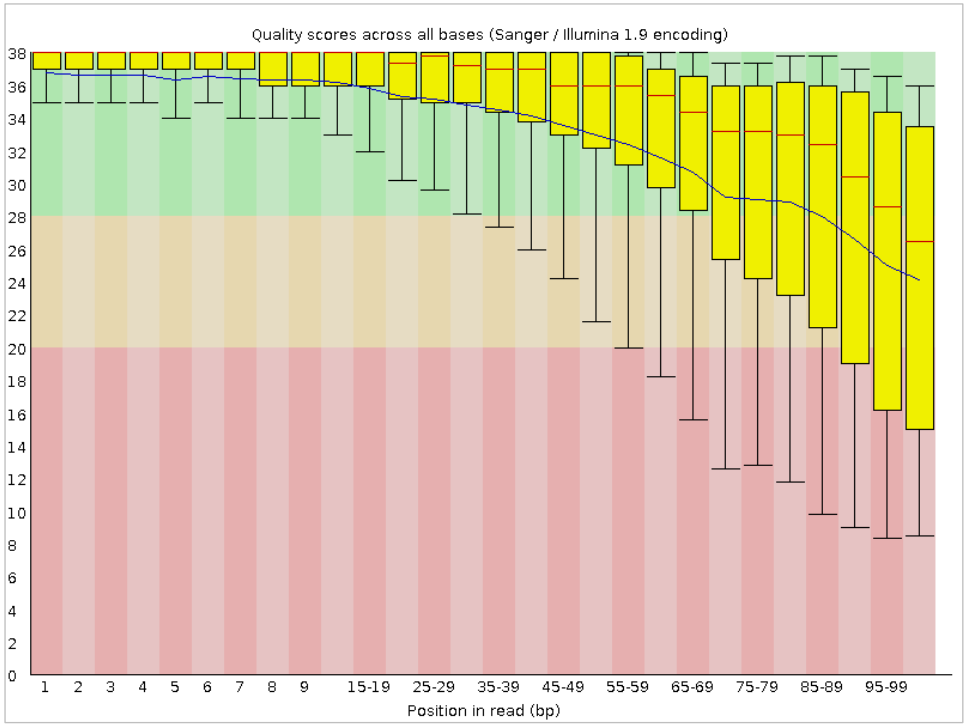
Agenda

1. Background and Definitions
2. Why You Should Care About Your Upstream?
3. **A Drink from the Bioinformatics Firehose**
   - File formats
   - Popular tools
   - QA Filtering
   - Visualization
4. Service Provider Deliverables: CEPH T
5. Applications That Require Special Ups

FASTQ

- **Contains 3 things per read:**
  - Sequence identifier (unique)
  - Sequence bases [len N]
  - Base quality scores [len N]
- **Often “gzip” compressed (fq.gz)**
- **If not demultiplexed, first 4 or 6bp is the “barcode” index.** Used to split lanes out by sample.
- **Filtering may include:**
  - Removing adapters & primers
  - Clip poor quality bases at ends
  - Remove flagged low-quality reads

```
@HWI-ST845:4:1101:16436:2254#0/1
CAAACAGGCATGCGAGGTGCCTTTGGAAAGCCCCAGGGCACTGTGGCCAG
+
Y\[SQORPMPYR\SNP_\][_babBBBBBBBBBBBBBBBBBBBBBBBBB
```



SAM/BAM

- **Spec defined by samtools author** Heng Li, aka Li H, aka lh3.
- **SAM is text version** (easy for any program to output)
- **BAM is binary/compressed version** with indexing support
- **Alignment in terms of code of matches, insertions, deletions, gaps and clipping**
- **Can have any custom flags set** by analysis program (and many do)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Key Fields

- Chr, position
- Mapping quality
- CIGAR
- Name/position of mate
- Total template length
- Sequence
- Quality

VCF

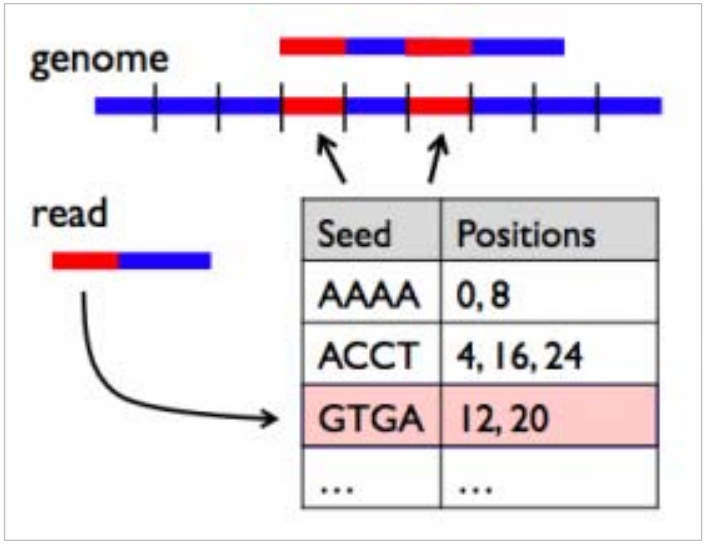
- **Specification defined by the 1000 genomes group (now v4.1)**
- **Commonly compressed indexed with bgzip/tabix** (allows for reading directly by a Genome Browser)
- **Contains arbitrary data per "site" (INFO fields) and per sample**
- **Single-Sample VCF:**
  - Contains only the variants for the sample.
- **Multi-Sample VCF:**
  - Whenever one sample has a variant, all samples get a "genotype" (often "ref")
- **Caveat:**
  - VCF requires a reference base be specified. Leaving insertions to be "encoded" 1bp differently than they are annotated

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Sa
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth"
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequer
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Al
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membershi
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 members
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have d
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype C
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype
```

1	2	3	4	5	6	7	8	9	10	11
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	?	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	?	T	?	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTCT	G,GTACT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

Aligners

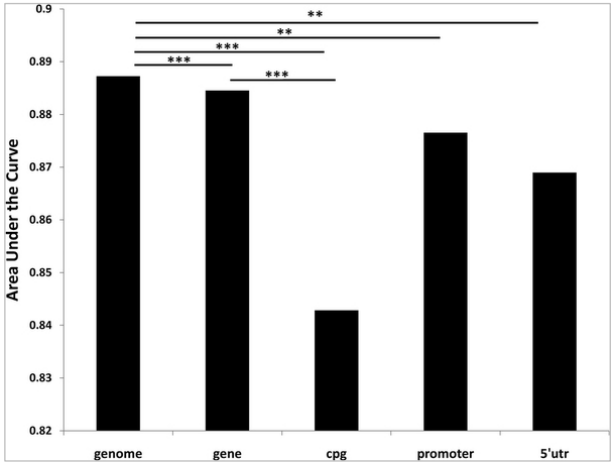
- **BWA (also by Li H)**
  - Most prolifically used for genome alignment
  - BWA-SW version geared for long reads (>100bp)
  - Supports aligning with insertions/deletions to reference
- **Bowtie** (John Hopkins, part of “Tuxedo” suite)
  - Very fast, used commonly in RNA-Seq workflows
  - Version 1 did not support “gapped” alignments
  - Bowtie2 supports local gapped, longer reads
- **Novoalign, Eland, SOAP, MAQ,**
  - Seed and expand strategy
- **TopHat, SHRiMP, STAR, Gmap**
  - Specifically designed for ESTs
- **Most improved by paired-end (mate-pairs)**



InDel Re-Alignment

- **Place, then realign “de Novo”**
  - Each read aligned independently by global aligner.
  - May have different preference of how to handle “gaps” to reference.
- **Local Re-Aligners for InDels**
  - Pindel
  - GATK
- **Important areas still problematic:**
  - CpG islands
  - Promoter and 5′-UTR regions of the genome

```
reference CAATC    realignment CAATC
read1     CA-TC       ---->    CA-TC
read2     C-ATC                CA-TC
```



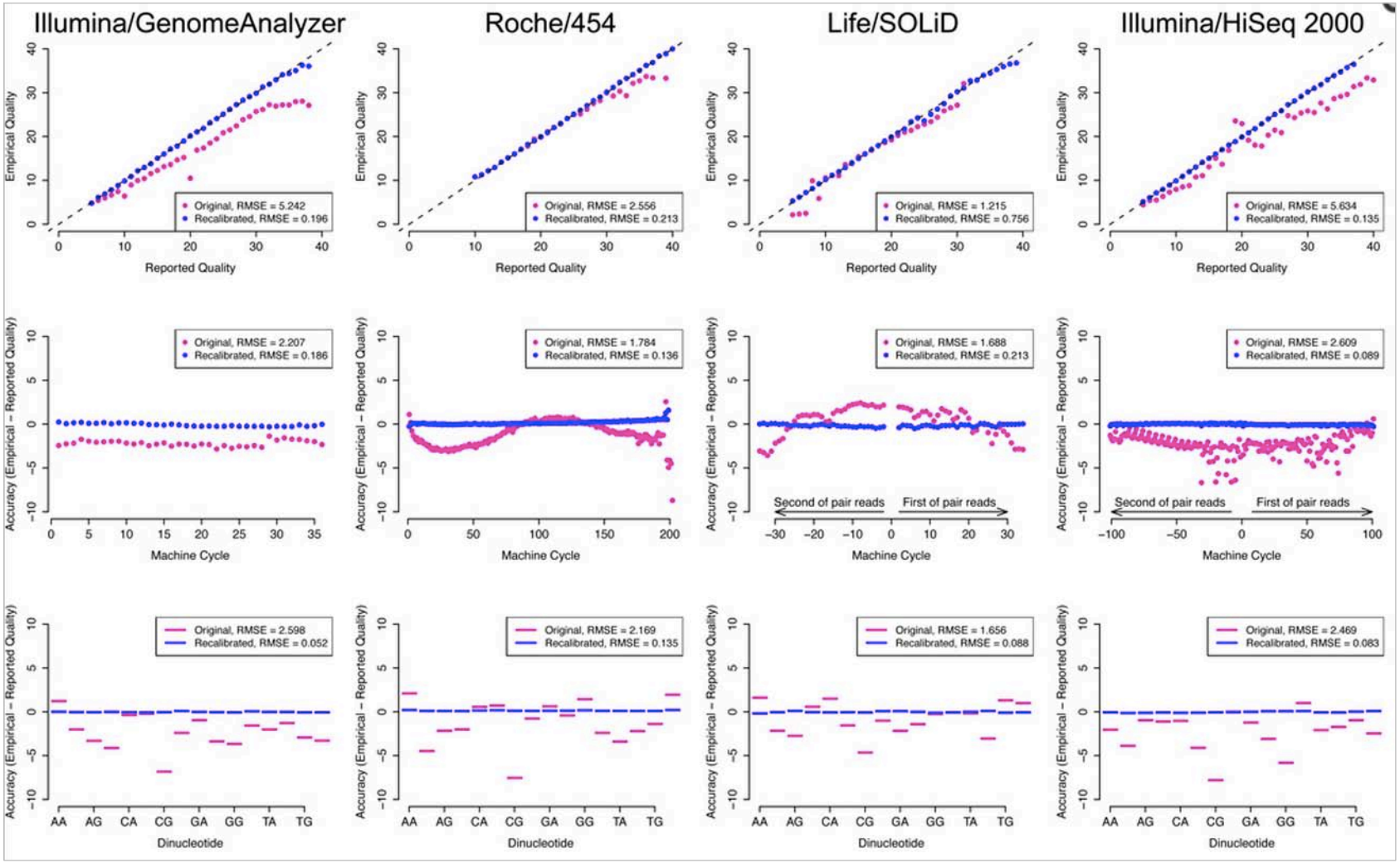
AUC (area under the curve) comparison for different genetic regions.

Weixin Wang. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Scientific Reports 1, Article number: 55

Variant Callers

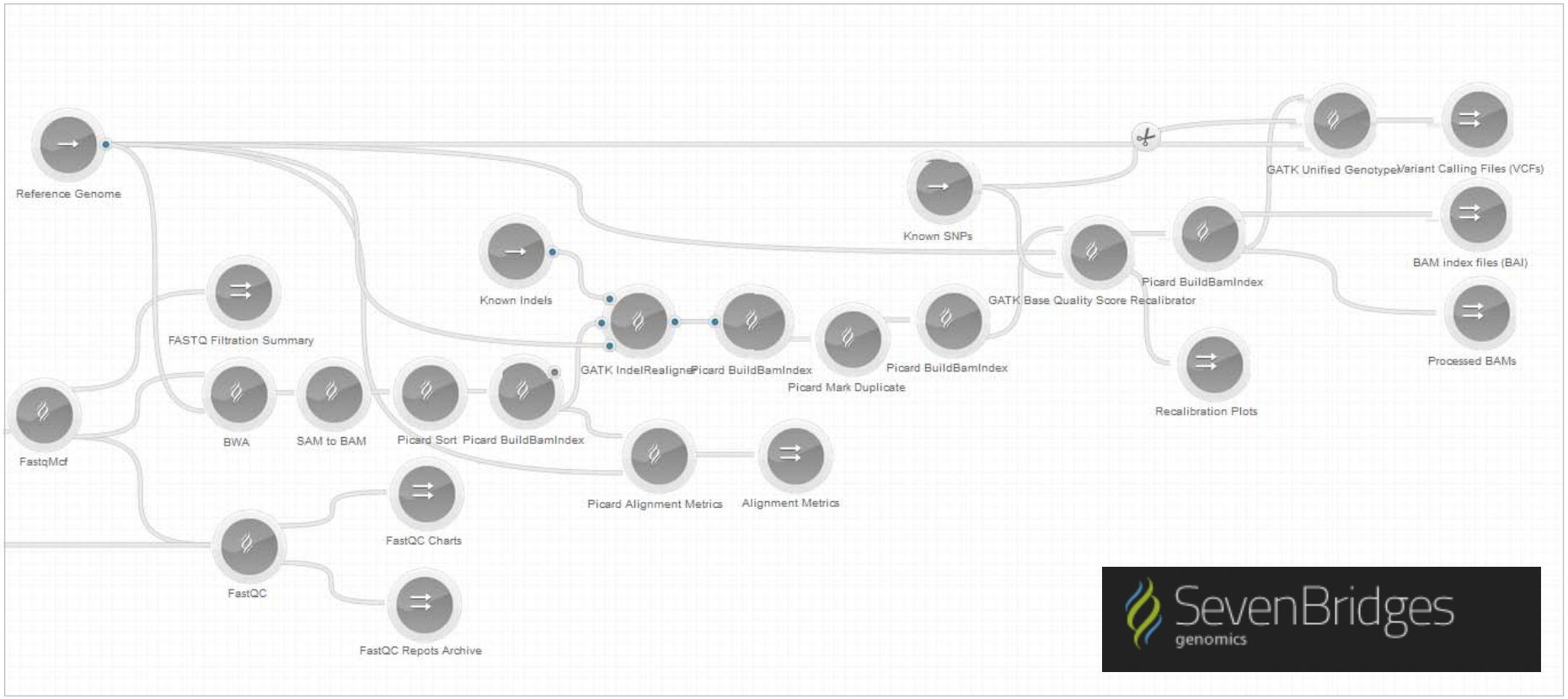
- **Samtools**
  - “mpileup” command computes BAQ, preforms local realignment
  - Many filters can be applied to get high-quality variants
- **GATK**
  - More than just a variant caller, but UnifiedGenotyper is widely used
  - Also provides pre-calling tools like local InDel realignment and quality score recalibration
- **Custom tools specific to platform:**
  - CASAVA includes a variant caller for illumana whole-genome data
  - Ion Torrent has a caller that handles InDels better for their tech

Quality Score Recalibration



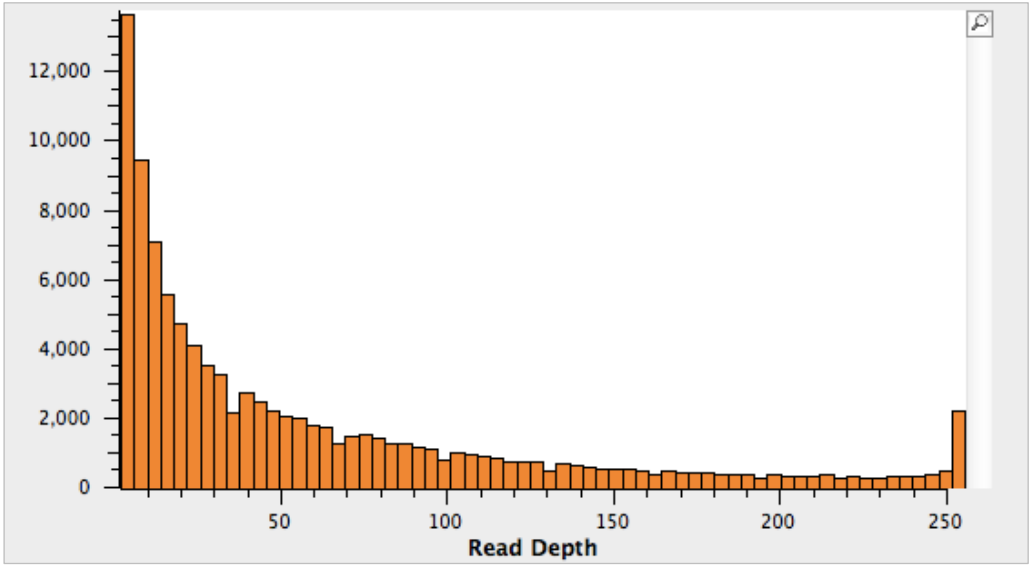
DePristo (2011) A framework for variation discovery …. Nature Genetics 43(5) 491

BWA+GATK Best Practices Pipeline



Getting to High Confidence Variants

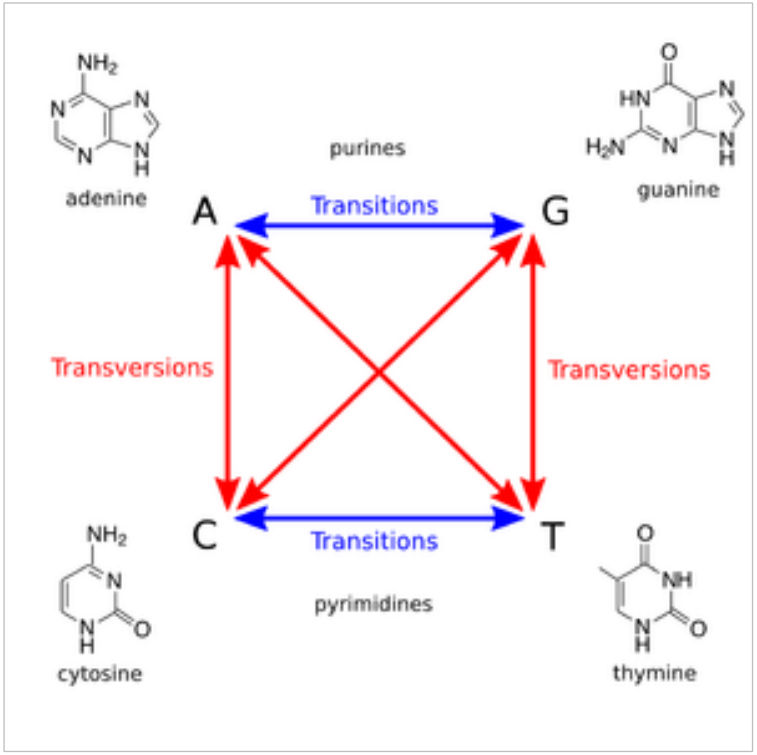
- Hard filters versus heuristic based statistics
- <10bp considered threshold for “low coverage”
- Quality score recalibration



		Unfiltered	Provided	RD>10 & GQ>20	Exonic
Gabe	**SNPs**	98621	89132	65009	19365
	InDels	8141	7800	6503	428
	Ts/Tv	2.36	2.45	2.54	3.26
Trio	**Mendel Errors**	234	202	46	3

Ti/Tv

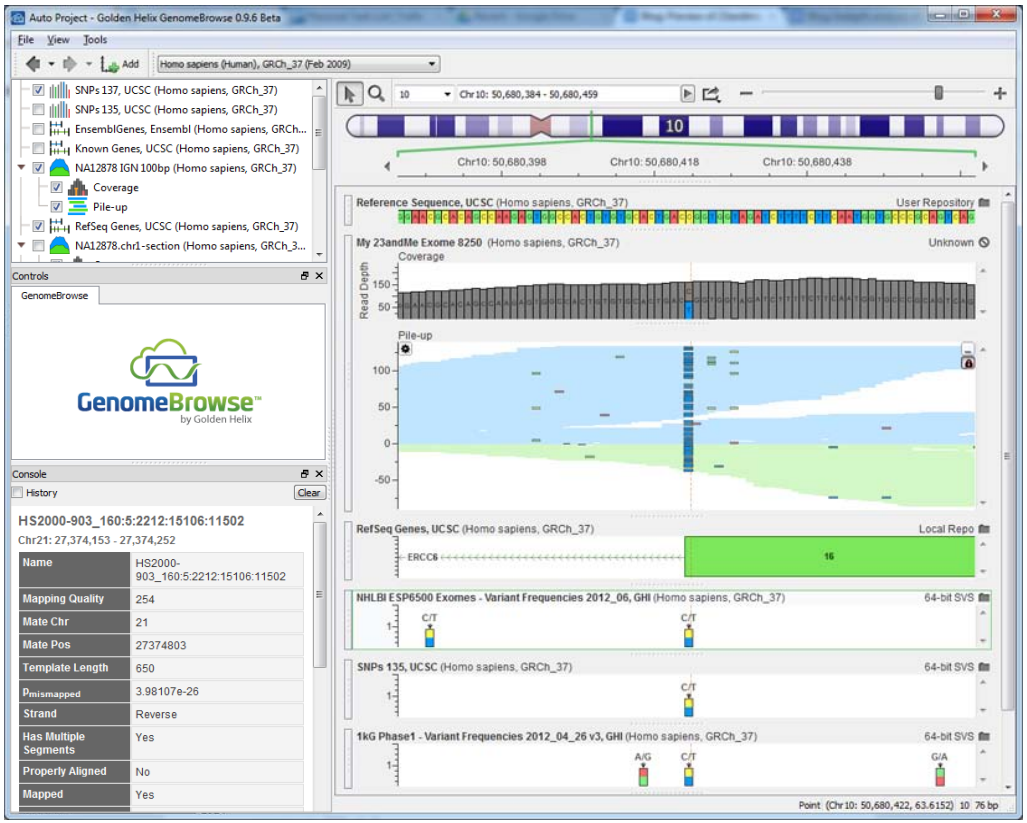
- **Ts/Tv ratio can measure true biological ratio of mutation types versus sequence error:**
  - Random seq errors: 2/4 or 0.5
  - Genome-wide: ~2.0-2.1
  - Exome capture: ~2.5-2.8
  - Coding: ~3.0-3.3
- **Divergent too far than this indicates random sequence errors biasing the number.**



DePristo (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43(5) 491

Visualization

- **Genome browsers:**
  - Validate variant calls
  - Look at gene annotations, problematic regions, population catalogs
  - Compare samples where no variant called
- **Free Genome Browsers:**
  - **IGV**
    - Popular desktop by Broad
  - **UCSC**
    - Web-based, most extensive annotations
  - **GenomeBrowse**
    - Designed to be publication ready
    - Smooth zoom and navigation

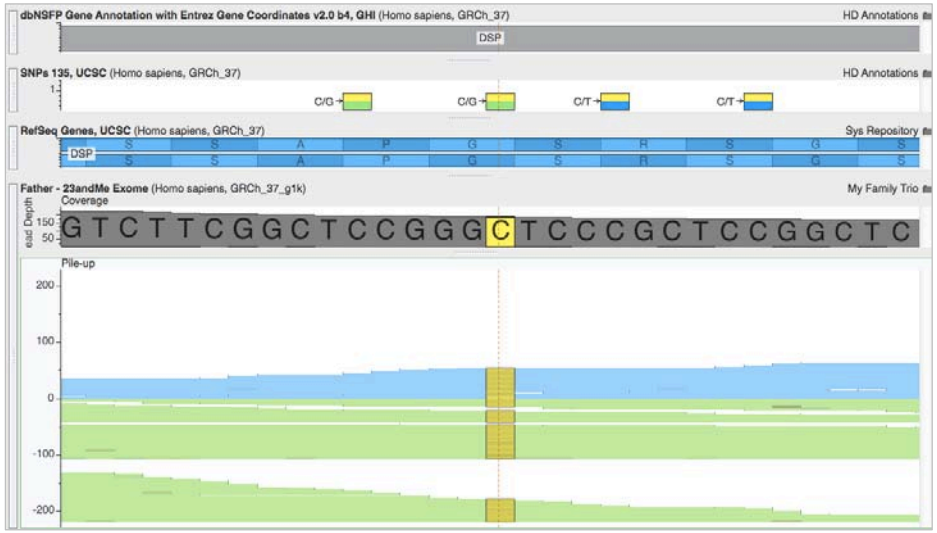


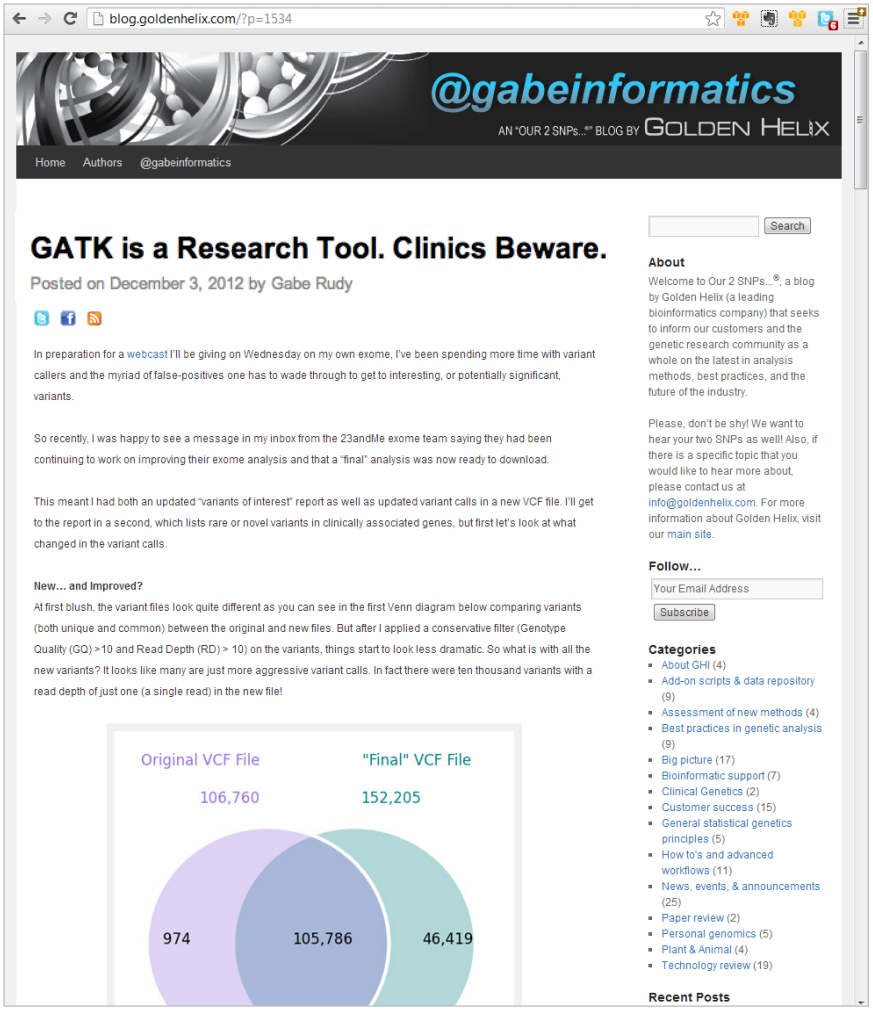
Updates in Software Can Introduce Bugs

- Found **8K phantom variants** in my "final" 23AndMe exome

List of selected variants

Variant 1:	**Gene:** DSP **Your genotype:** GC/GC **Location:** chr6:7585967	
Effect:	FRAME SHIFT	**Type:**HIGH
Frequency:	**1KGenomes:** NA	**dbSNP:** NA
Quality:	**Genotype quality:** 99.00	**Coverage depth:** 153
Details:	**Gene description:** desmoplakin **Transcript:** ENST00000379802 **EntrezId:** 1832 **UniProt:** P15924	 **AA change:** NA **EnsemblId:** ENSG00000096696 **OMIM:** 125647





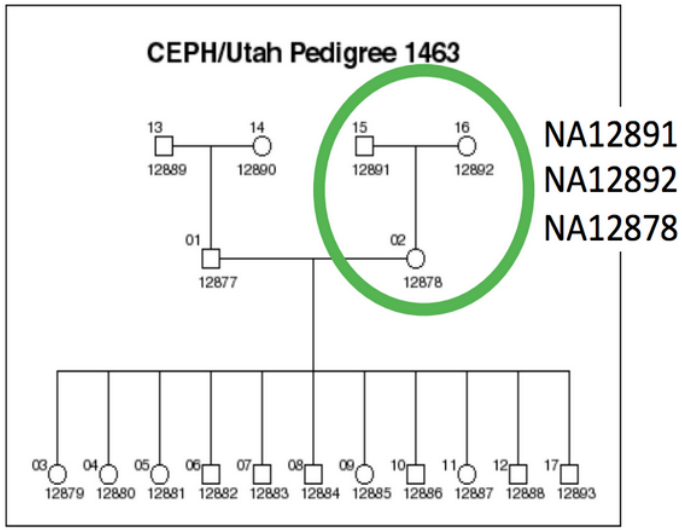
GOLDEN HELIX
Accelerating the Quest for Significance™

[My 23andMe “Buggy” Variant and Interpretation Example]

Agenda

1. Background and Definitions
2. Why You Should Care About Your Upstream?
3. A Drink from the Bioinformatics Firehose
4. **Service Provider Deliverables: CEPH Trio Example**
5. Applications That Require Special Upstream Analysis

Recent Blog Post



The State of NGS Variant Calling: Don't Panic!!

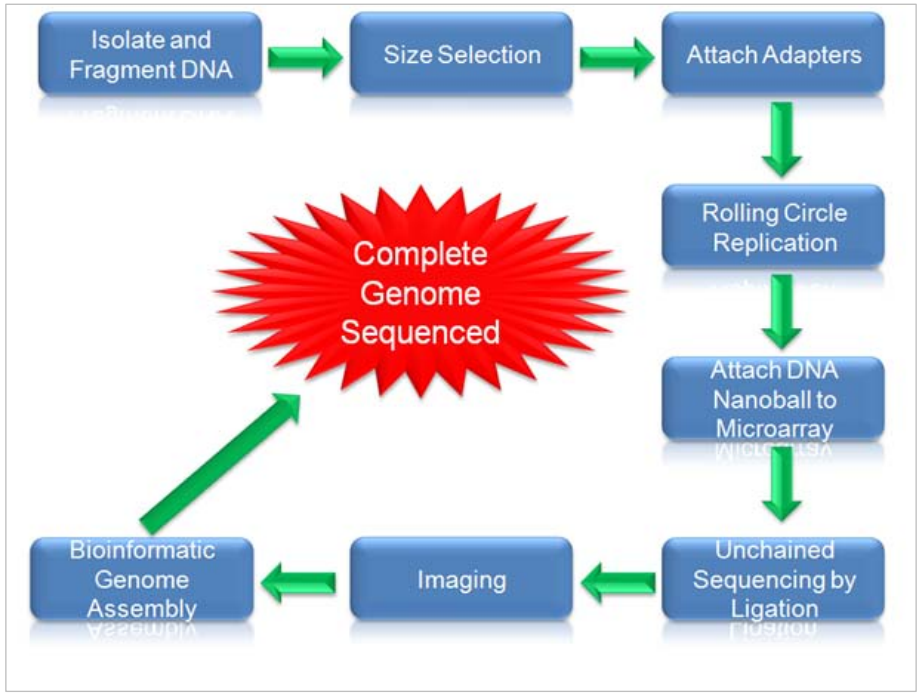
http://blog.goldenhelix.com/?p=1725

POLL:

From which sequencing provider(s) do you receive your upstream data from?

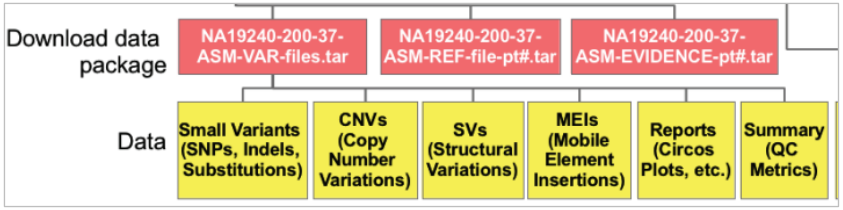
Complete Genomics

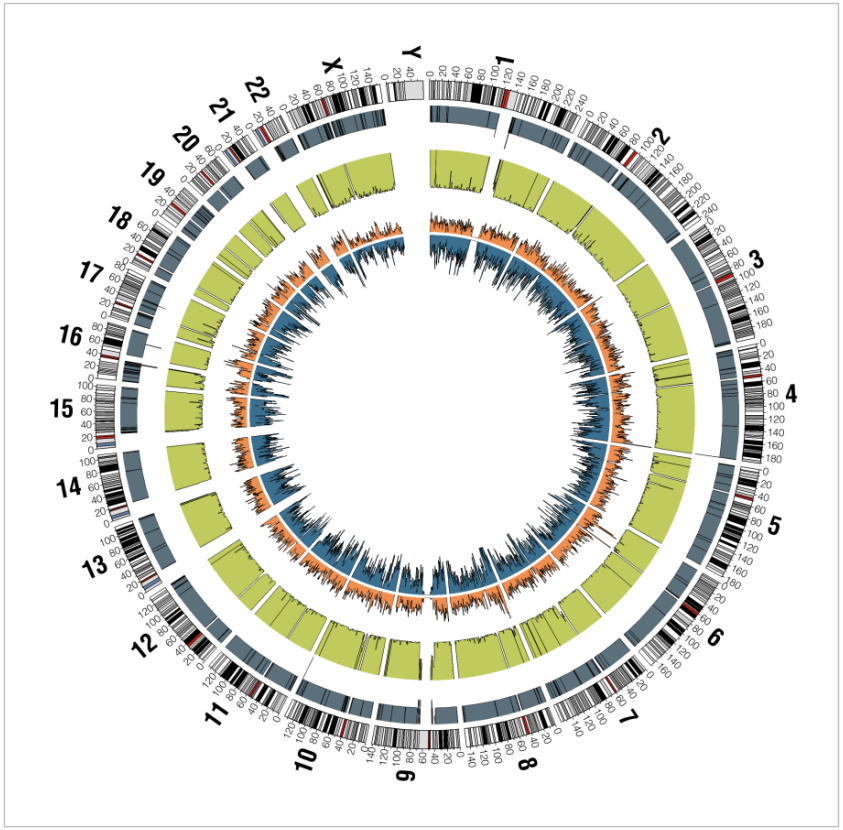
- **Different:**
  - Sequencing technology
  - Alignment/variant calling algorithms
  - File formats
- **But high quality:**
  - MNPs, Indels
  - CNV/SV calls
- **Whole genome only**
- **Also provide tumor/normal pair analysis**
- Being acquired by BGI**, some question their sustainability**



Complete Genomics Deliverables

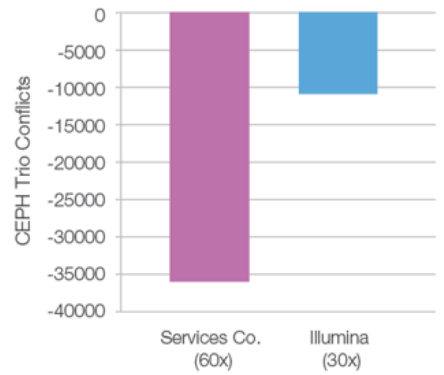
- **Summary statistics**
- **“var” and “masterVar” files**
  - Can be converted to VCF
  - Some tools (like SVS) can import them directly
- **Evidence files**
  - Can be converted to BAM
- **CNV, SV calls in text files**
  - CNV: Chr1:85980000-86006000 2.06 4x gain, covers DDAH1
  - SV: Chr21:27374158-27374699 common inversion



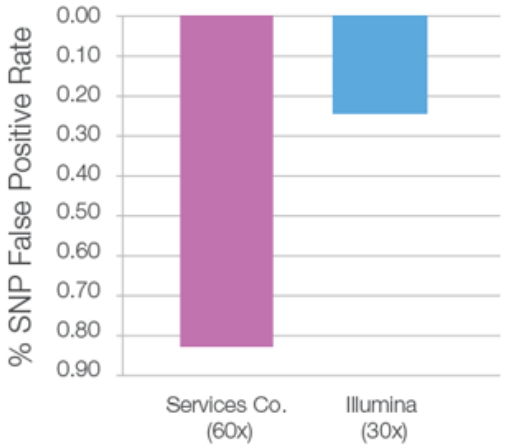


Illumina Genome Network

- **Standardized sequencing and analysis, but multiple labs may be contracted service provider.**
- **30x whole genomes**
  - SNPs, InDels, CNVs, SVs
  - Concordance with SNP array (provided)
  - Summary report
- **Illumina provided tools used**
  - CASAVA toolkit with ELAND aligner
- **Also provide Tumor/Normal pair**
  - Somatic SNVs and InDels identified by looking at the tumor/normal together



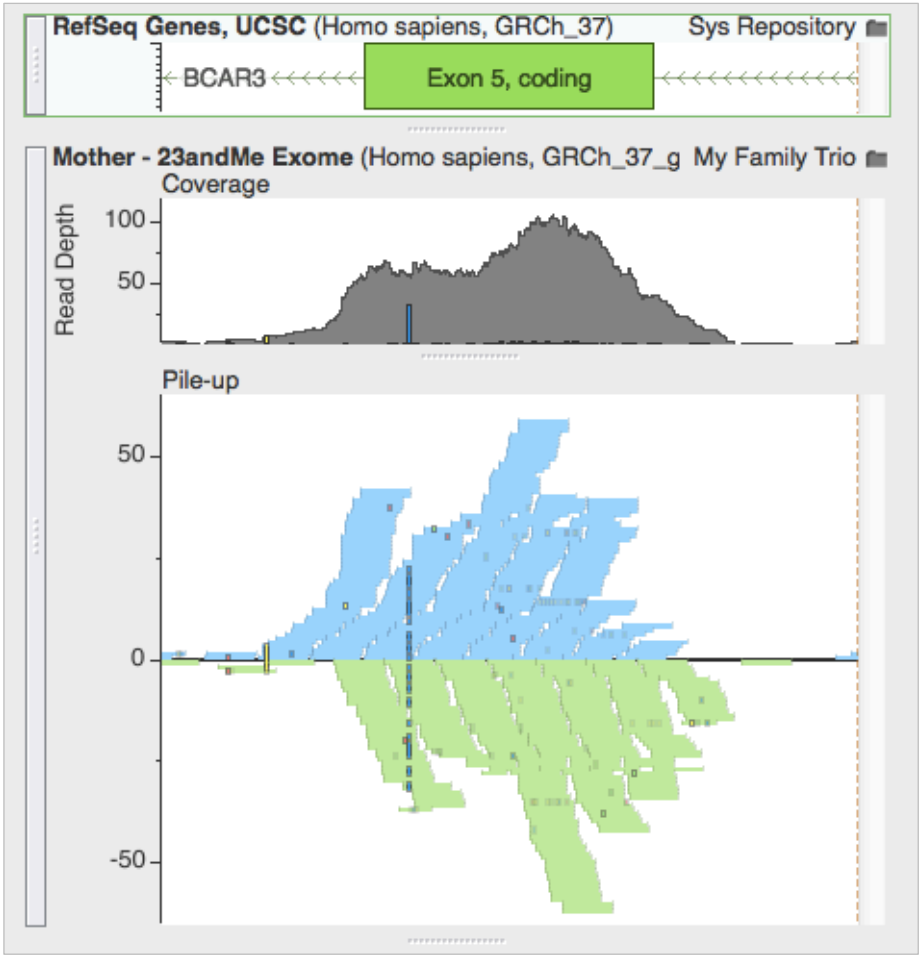
	SERVICES CO. (60X)	ILLUMINA (30X)
Callable Sites Present in Trio (bp) "N-Excluded"	2,575,974,378	2,530,389,955
Mendelian Conflicts on Genotypes	35,843	10,801
Conflicts Per Mb	13.91	4.27



	SERVICES CO. (60X)	ILLUMINA (30X)
Called SNPs in Any One of Trio	4,294,576	4,403,335
Mendelian Conflicts on Genotypes	35,843	10,801
False Positive Rate	0.83%	0.25%

Your Local Core Lab – Or 23andMe Exome Pilot!

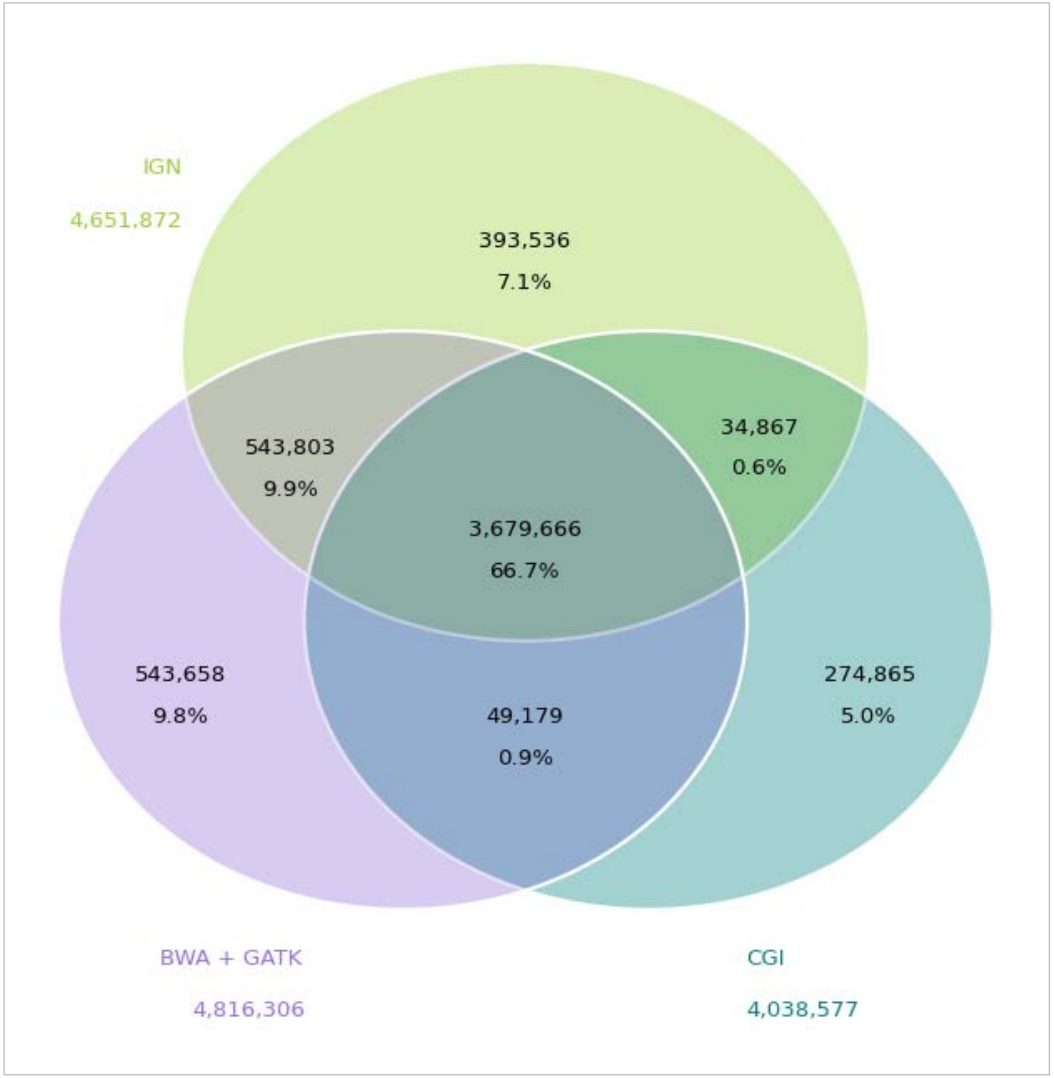
- **Research core labs often use a BWA+GATK pipeline**
  - Especially for exomes
- **Deliverables:**
  - VCF with SNVs, InDels
  - BAM
- **Tools for CNV/SV calling less standardized**
  - Not commonly attempted with exomes



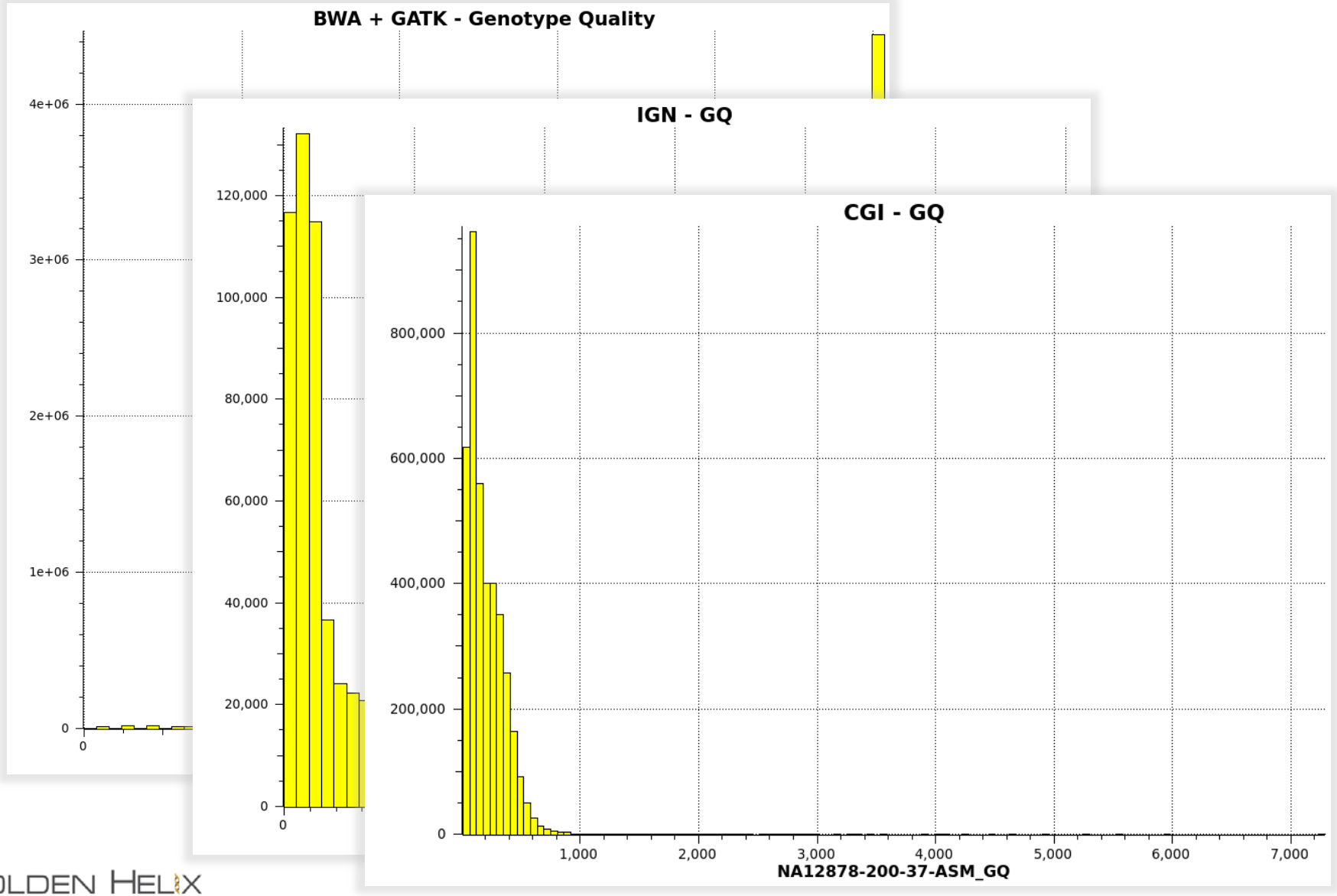
CEPH Trio

- **The “benchmark” trio.**
  - Child, female NA12878 may be the most sequenced cell line
  - Father NA12891 and Mother NA12892
- **Whole Genome Data for Trio**
  - CGI with v2 pipeline
  - IGN WGS at 30x, 100bp PE
  - “Core Lab” BWA + GATK Best Practices on 100bp PE
- **Concordance and Comparisons**
  - Lets interactively review examples where these three service providers differ and how.

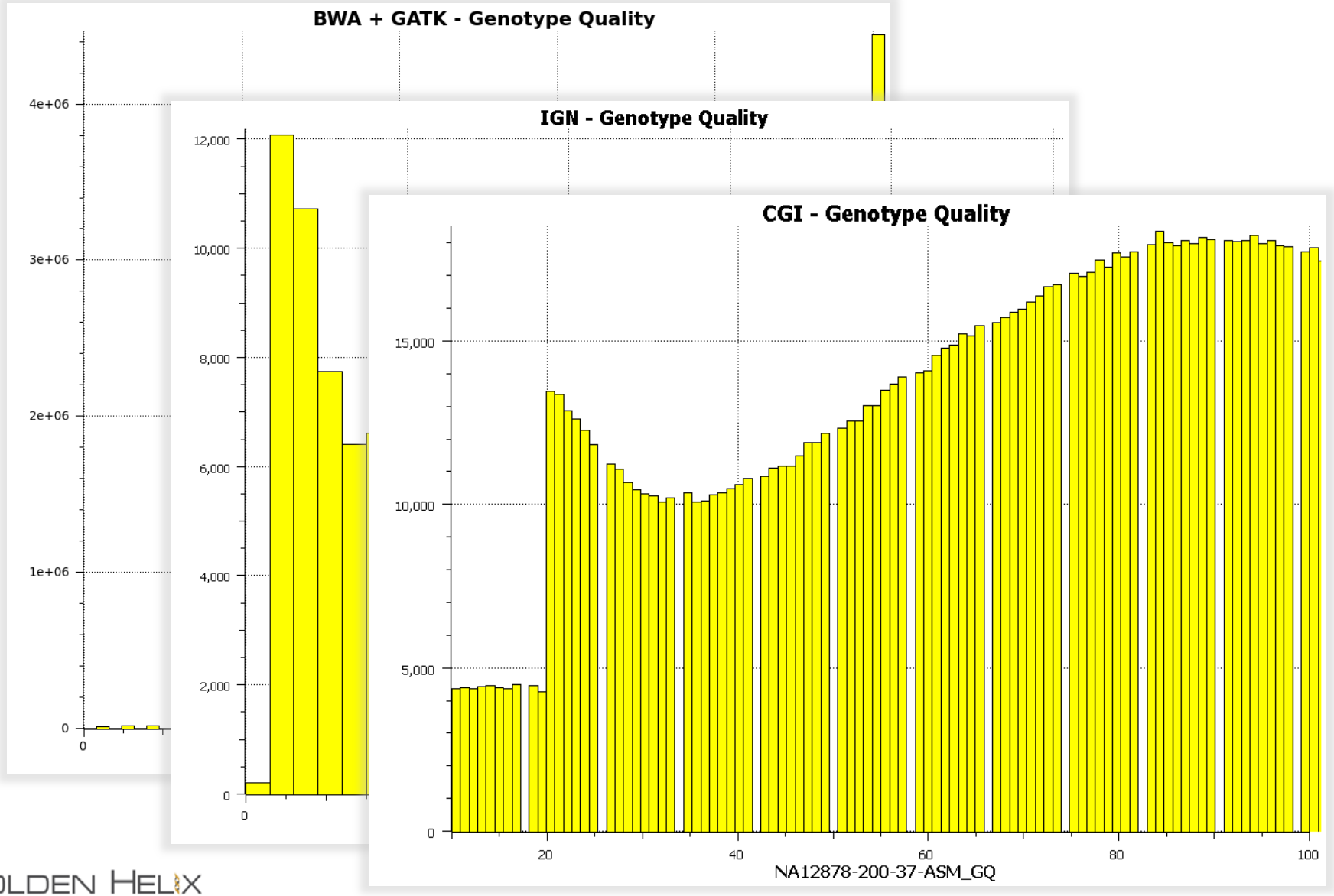
Total Imported Variants



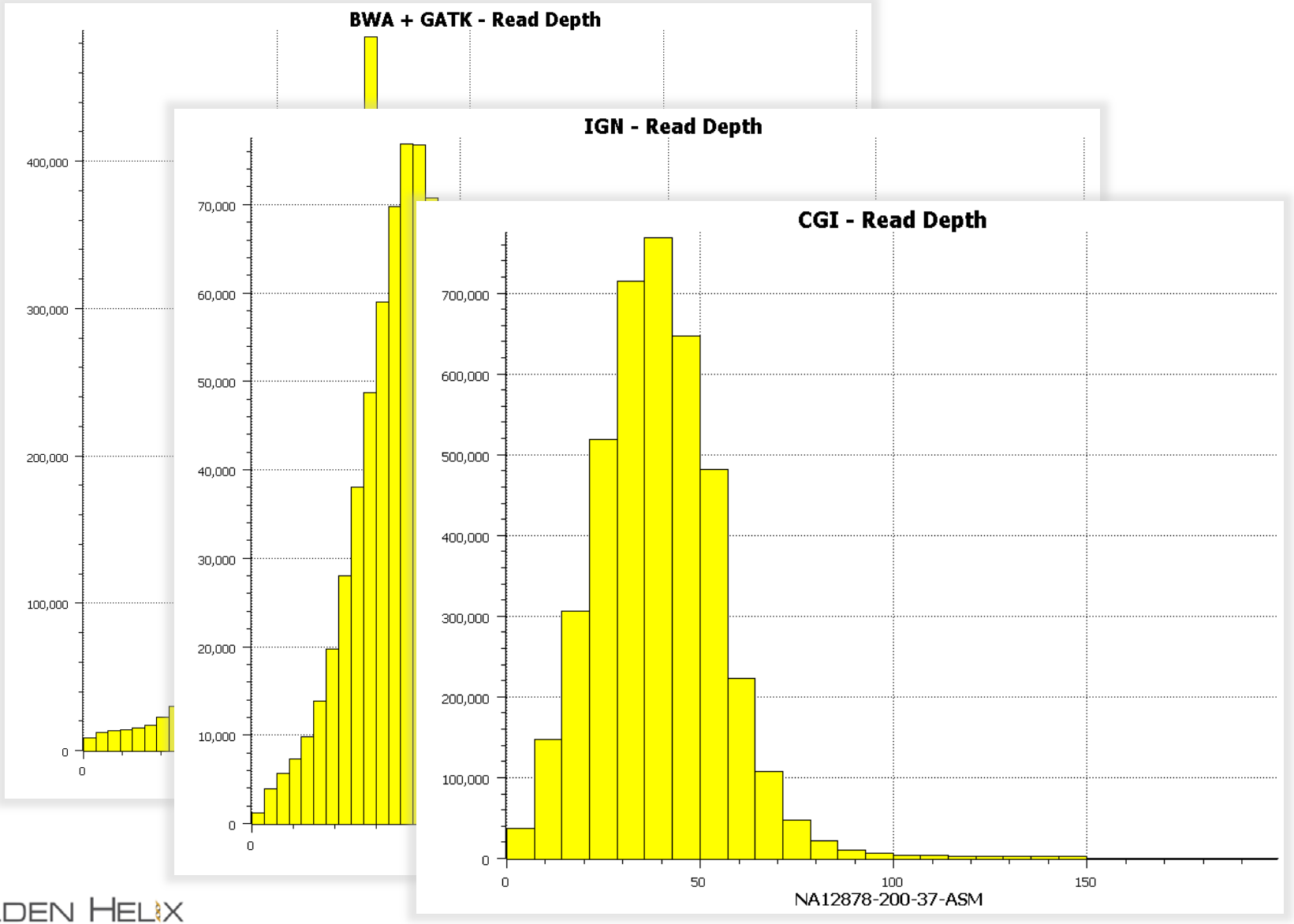
Genotype Quality



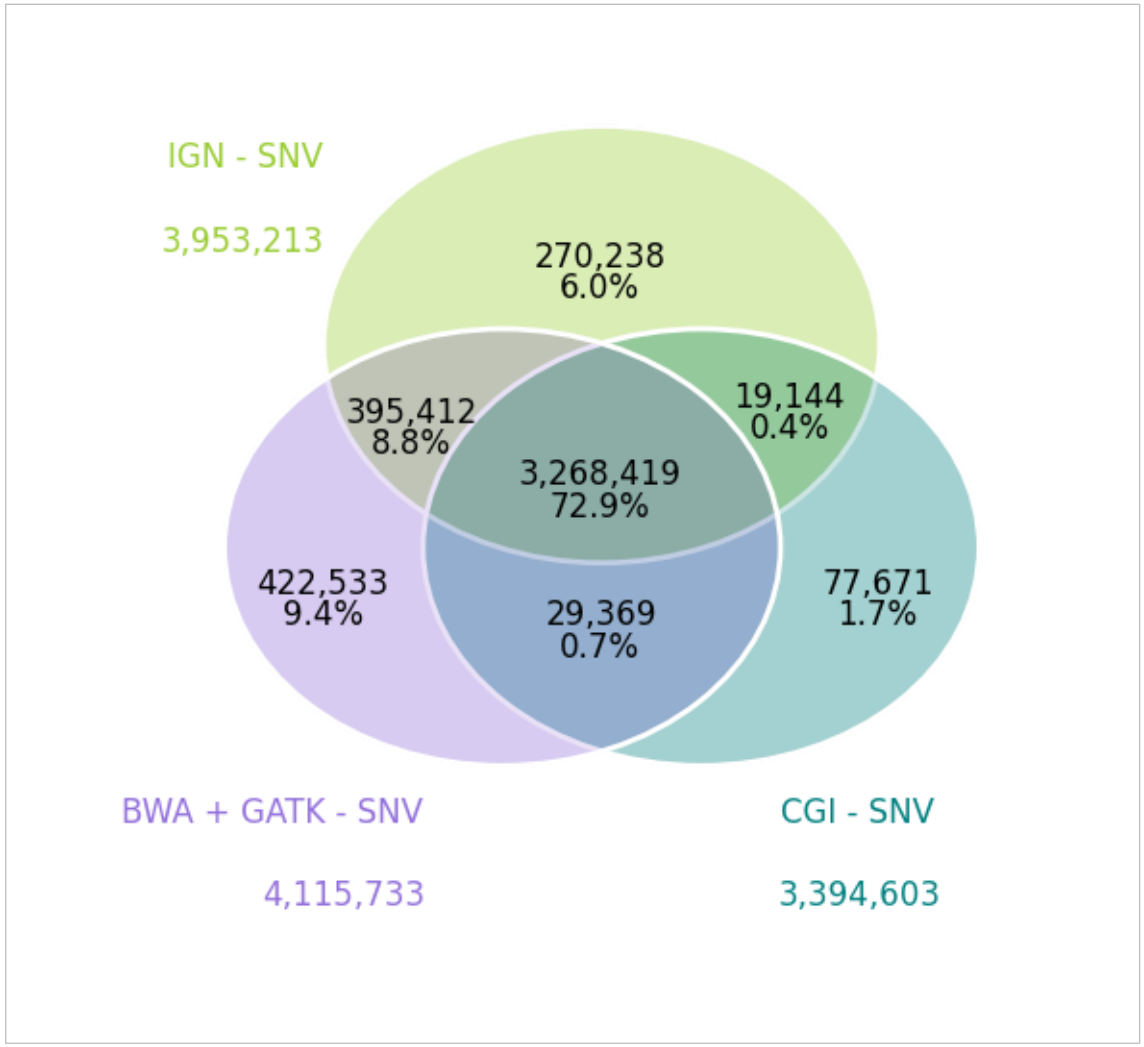
Genotype Quality



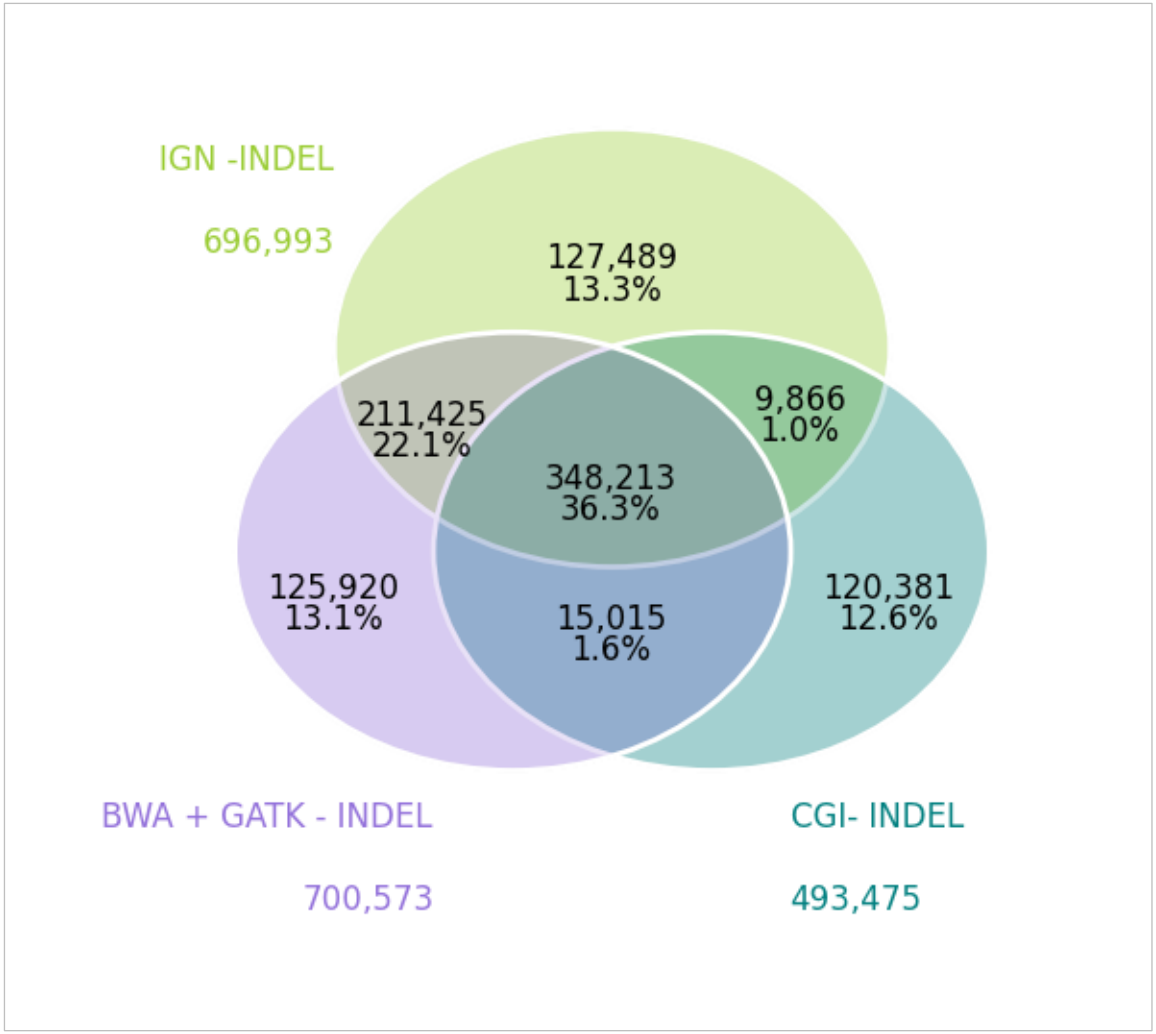
Read Depth


BWA + GATK - Read Depth
400,000
300,000
200,000
100,000
0
0
IGN - Read Depth
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
0
CGI - Read Depth
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
0
50
100
150
NA12878-200-37-ASM

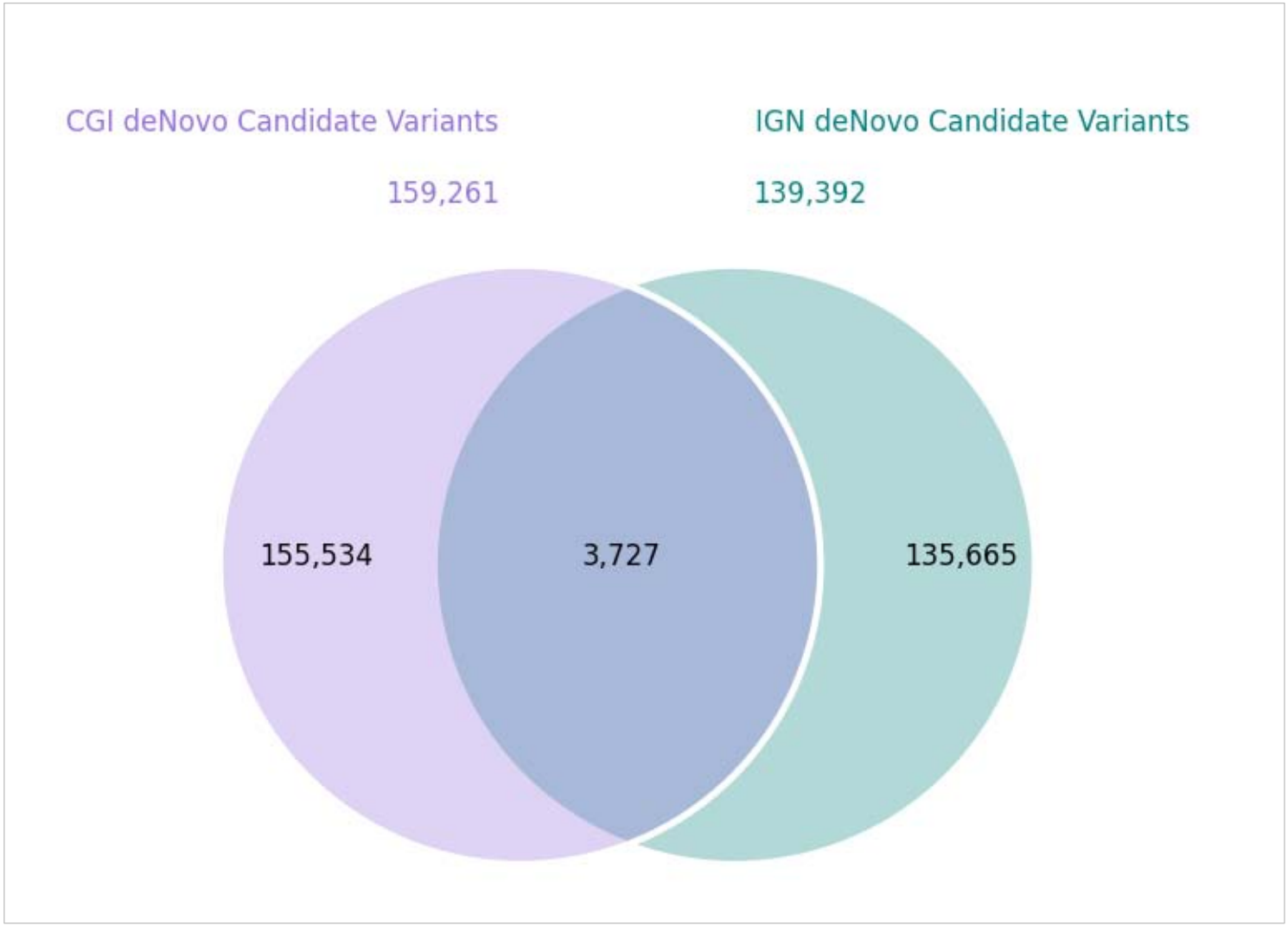
SNV Concordance Rate



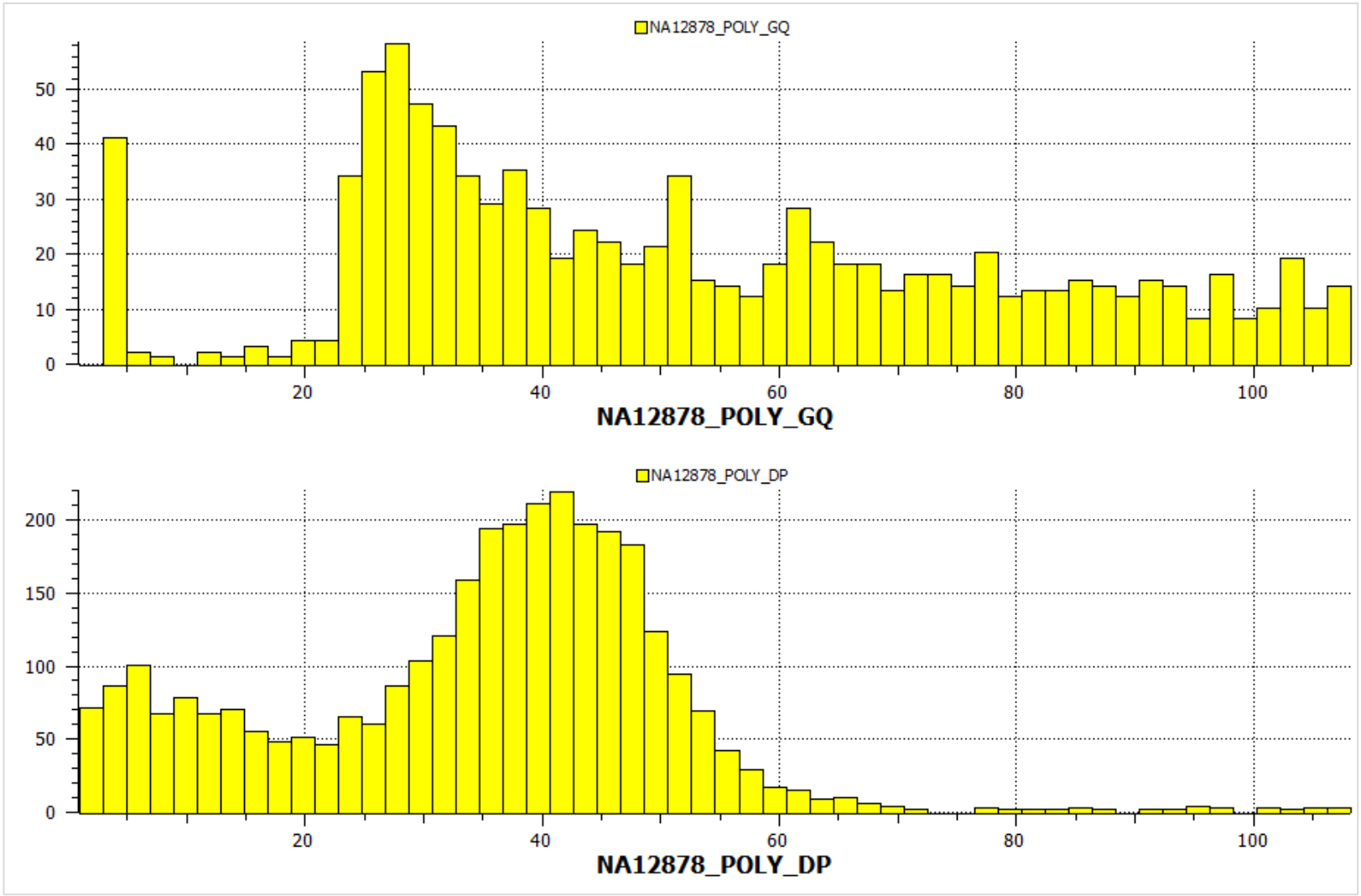
InDel Concordance Rate



De Novo Mutations in Trio



GQ and DP of Shared de Novo Mutations



[deNovo and SV/CNV of NA12878 trio]

Agenda

1. Background and Definitions
2. Why You Should Care About Your Upstream?
3. A Drink from the Bioinformatics Firehose
4. Service Provider Deliverables: CEPH Trio Example
5. **Applications That Require Special Upstream Analysis**

Applications That Require Special Upstream Analysis

- MHC Region
- Somatic Variant Calling
- RNA-Seq
- Alu and other repeats
- Phased variants and complex MNP
- Moving to a new reference genome

Somatic Sniper and Friends

- **Complete Genomics and IGN provide secondary alignment specific to tumor/normal pairs.**
- **Do variant calling with on BAMs on pair in conjunction**
- **SomaticSniper approach:**
  - Covered by at least 3 reads
  - Consensus quality of at least 20
  - Called a SNP in the tumor sample with SNP quality of at least 20
  - Maximum mapping quality of at least 40
  - No high-quality predicted indel within 10 bp
  - No more than 2 other SNVs called within 10 bp
  - Not in dbSNP (non-cancer dbSNP)
  - LOH filter (germline is het and tumor is homozygous)

What's Next?

The Analysis and Interpretation of My DTC 23andMe Exome

Short Courses

Assembly and Alignment

blog.goldenhelix.com

Download *free* **GenomeBrowse™**

www.goldenhelix.com

Killer App

Questions?

Use the Questions pane in your GoToWebinar window