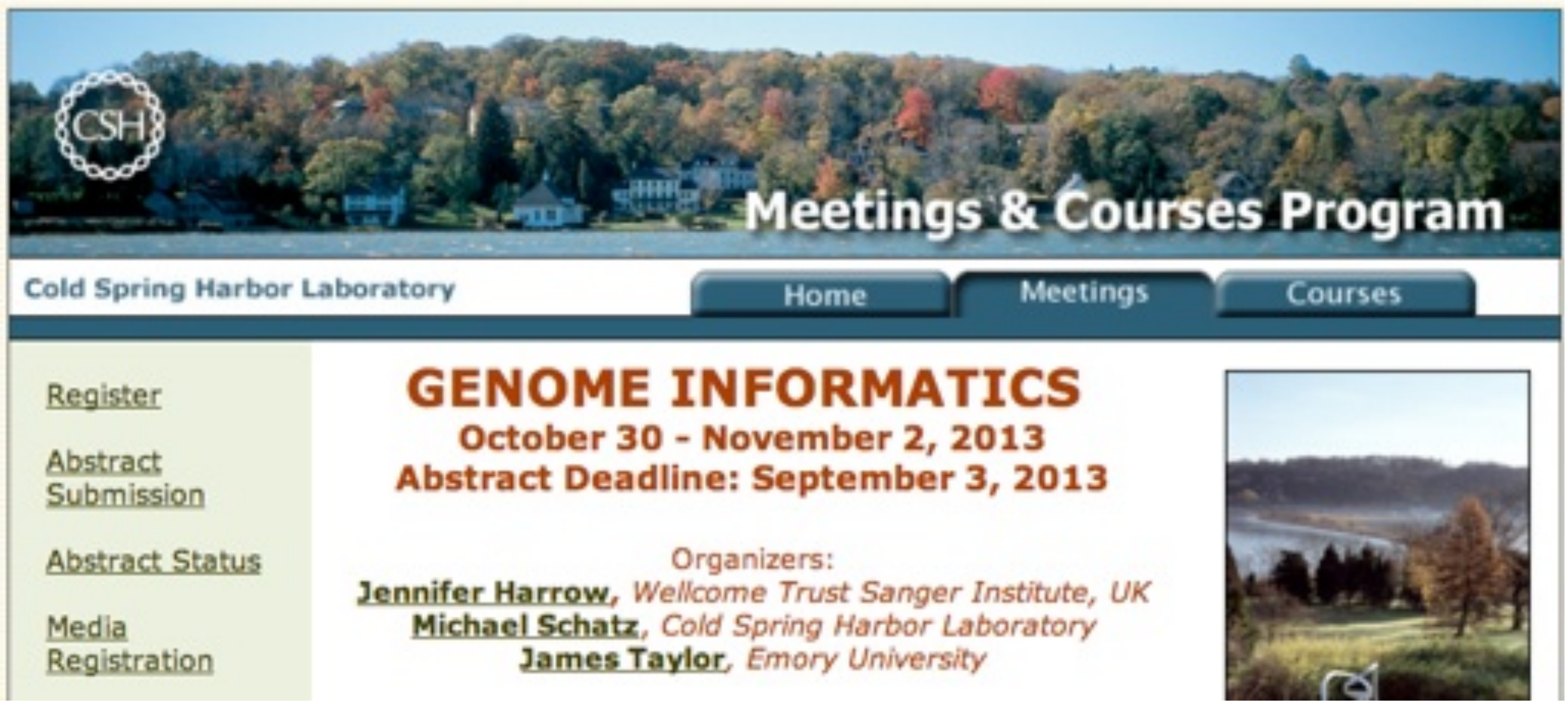
Lauri Saag

29. November 2013

Sequencing platform bias



GENOME INFORMATICS

October 30 - November 2, 2013

Abstract Deadline: September 3, 2013

Organizers:

Jennifer Harrow, *Wellcome Trust Sanger Institute, UK*

Michael Schatz, *Cold Spring Harbor Laboratory*

James Taylor, *Emory University*

A note on high coverage NGS SNP reproducibility

KEYNOTE SPEAKER I

Dr. Michael Snyder

Stanford University School of Medicine

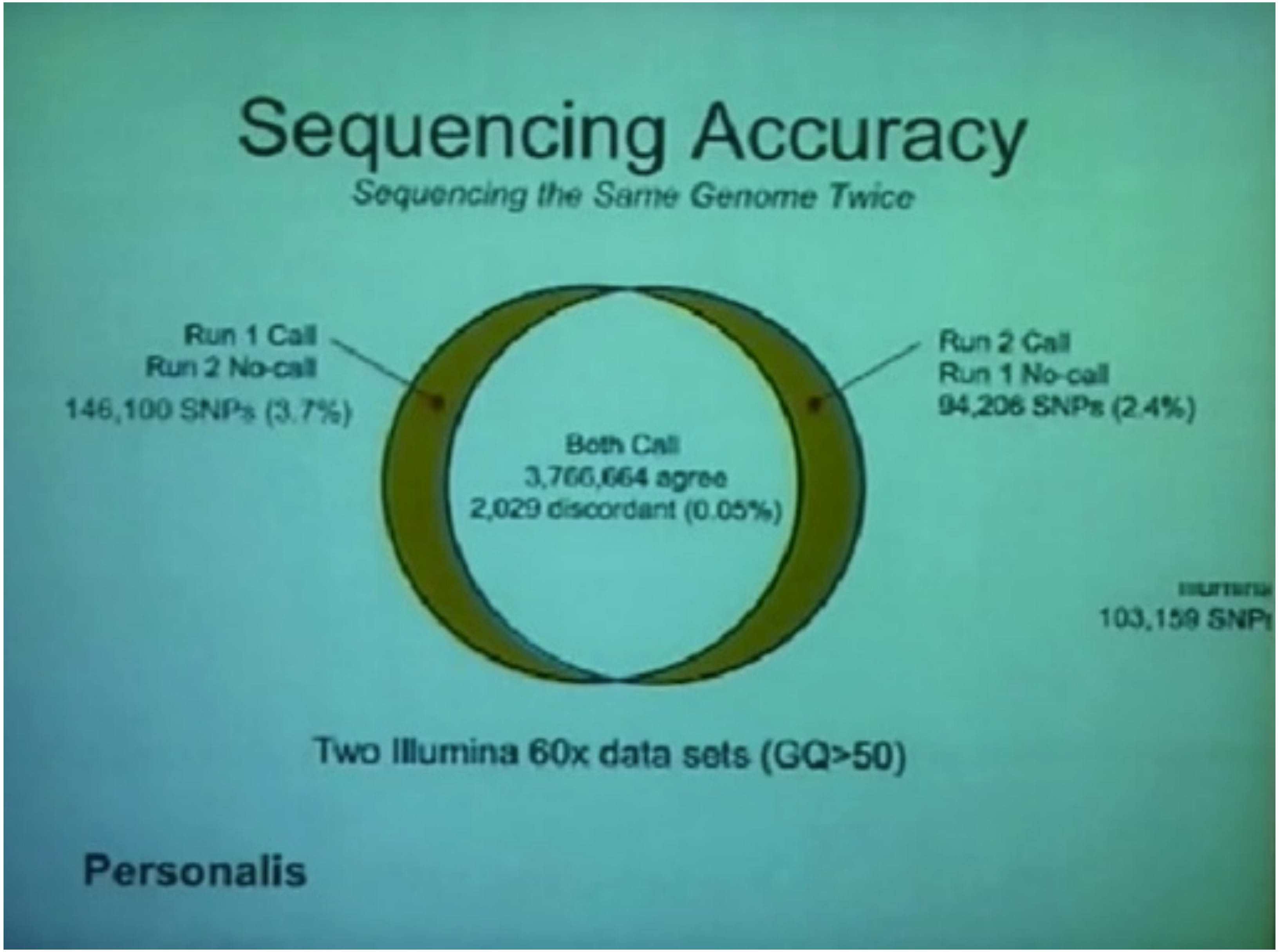
GENOME INFORMATICS

"Personal Omics Profiling: Monitoring Genomes, Transcriptomes and other Omes During Healthy and Disease States”

Dr. Michael Snyder

Stanford University School of Medicine

- Same sample sequenced twice with same methods
- Illumina, **60x coverage**
- Run 1: 3.948 M SNPs called
- Run 2: 3.925 M SNPs called
- 6% of total 4 M unique SNPs don’t overlap due to no-calls
- Of the overlapping SNPs, 0.05% are discordant



Sequencing Accuracy

Sequencing the Same Genome Twice

Run 1 Call
Run 2 No-call
146,100 SNPs (3.7%)

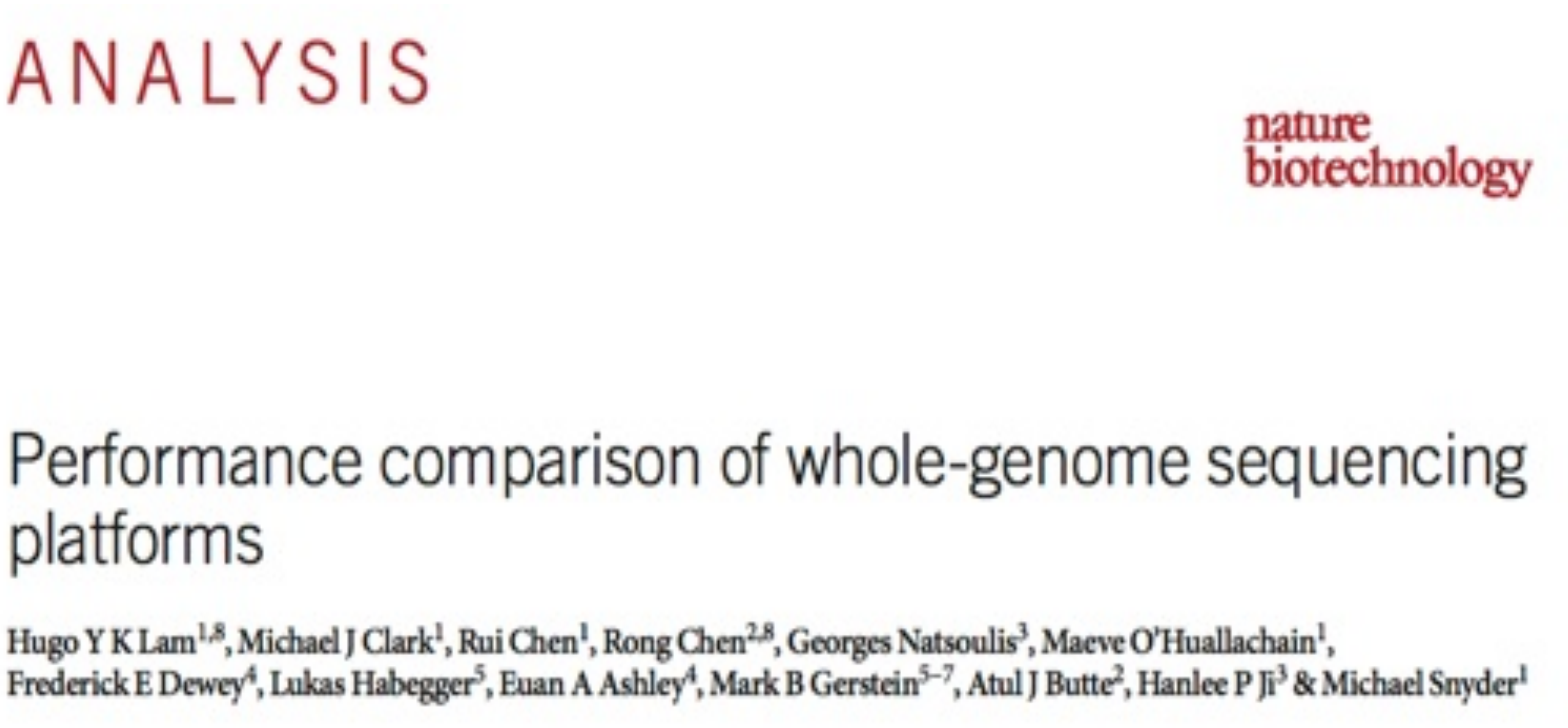
Both Call
3,766,664 agree
2,029 discordant (0.05%)

Run 2 Call
Run 1 No-call
94,206 SNPs (2.4%)

llumina
103,159 SNP

Two Illumina 60x data sets (GQ>50)

Personalis



ANALYSIS

nature biotechnology

Performance comparison of whole-genome sequencing platforms

Hugo Y K Lam[1,8], Michael J Clark[1], Rui Chen[1], Rong Chen[2,8], Georges Natsoulis[3], Maeve O'Huallachain[1], Frederick E Dewey[4], Lukas Habegger[5], Euan A Ashley[4], Mark B Gerstein[5–7], Atul J Butte[2], Hanlee P Ji[3] & Michael Snyder[1]

Complete Genomics vs. Illumina

Presentation emphasis on

overlap and concordance of SNVs

(single nucleotide variants)

- 1 individual sequenced on Illumina and CG platform
- 2 DNA sources: blood (mononuclear cells) and saliva

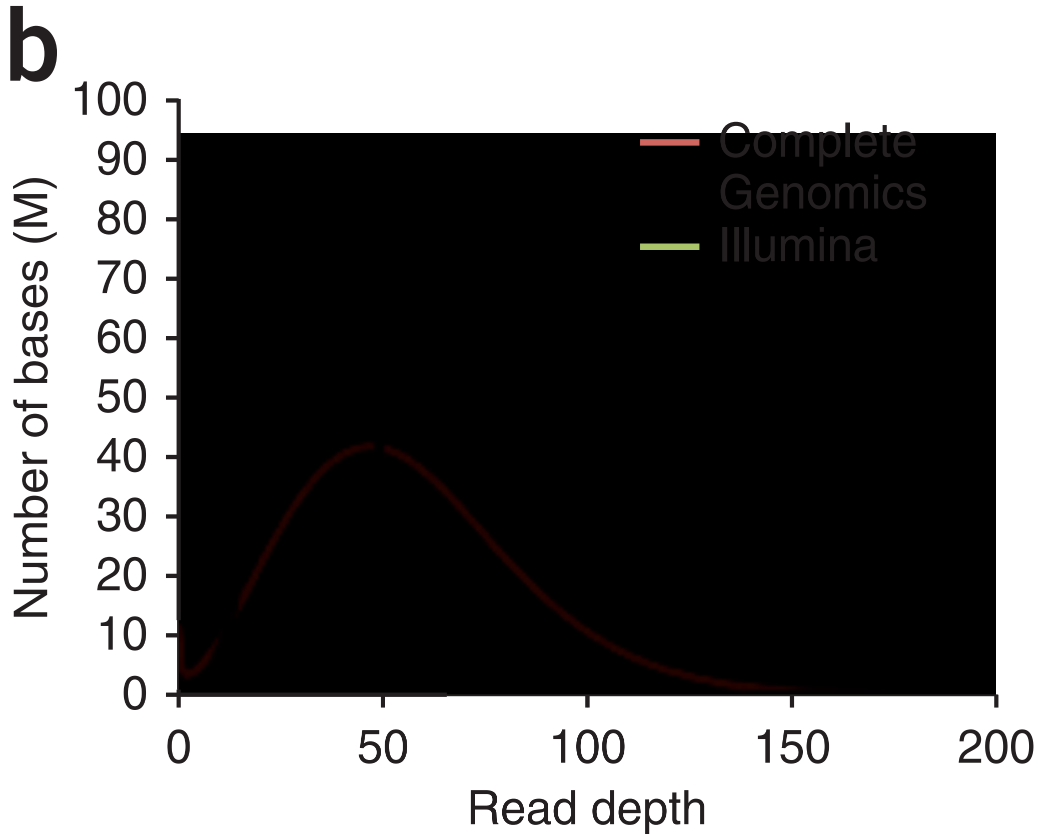
- Illumina HiSeq 2000: 101-bp paired-end reads
- CG: 35-bp paired-end reads
- Illumina
  - mapped with BWA
  - SNVs called with GATK
- CG mapped and called with CG pipeline

Lam et al. 2012

Lam et al. 2012

Average coverage 76x

CG less uniform in coverage



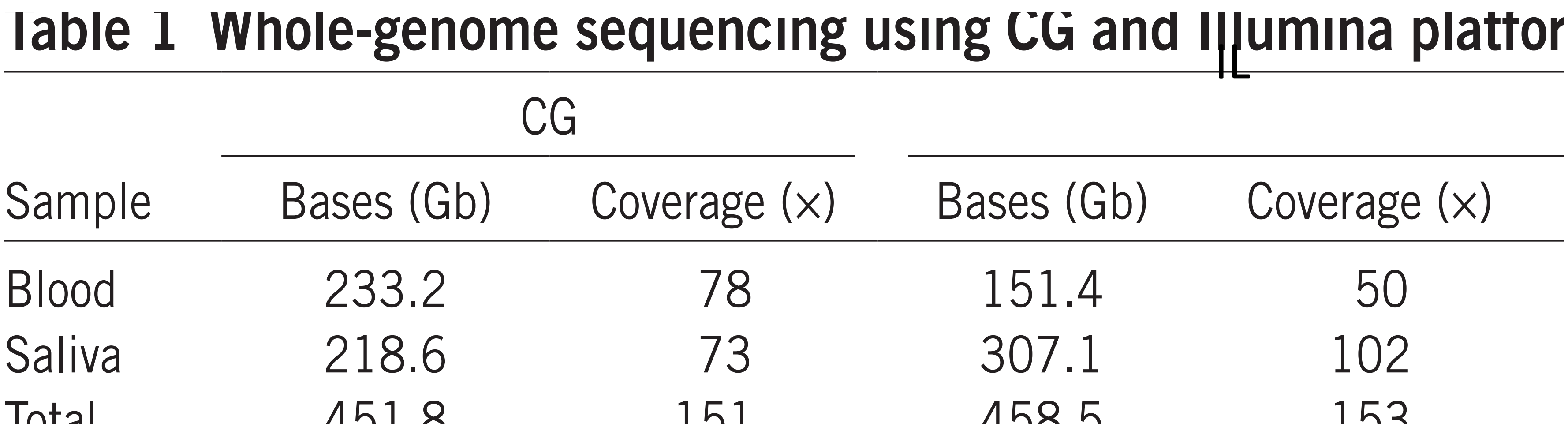


Table 1 Whole-genome sequencing using CG and Illumina platfor

	CG		IL	
Sample	Bases (Gb)	Coverage (×)	Bases (Gb)	Coverage (×)
Blood	233.2	78	151.4	50
Saliva	218.6	73	307.1	102
Total	451.8	151	458.5	153

- Data from blood and saliva combined
- SNVs detected from only one source discarded
- Only few of the tissue-specific calls could be validated by independent methods

Complete Genomics

Blood

Saliva

3,277,339

3,286,645

Illumina

Blood

Saliva

Lam et al. 2012

Independent validation

Omni Quad 1M Genotyping array

- Of 260,112 calls detected with Omni array, 99.5% present, 99.34% concordant, only 0.16% platform-specific SNVs.

Lam et al. 2012

Independent validation

Omni Quad 1M Genotyping array

- Of 260,112 calls detected with Omni array, 99.5% present, 99.34% concordant, only 0.16% platform-specific SNVs.

Sanger sequencing: randomly selected SNVs, both concordant and platform specific

- Validated 20 of 20 concordant SNVs; 2 of 15 (13.3%) Illumina-specific and 17 of 18 (94.4%) CG-specific

Lam et al. 2012

Independent validation

Omni Quad 1M Genotyping array

- Of 260,112 calls detected with Omni array, 99.5% present, 99.34% concordant, only 0.16% platform-specific SNVs.

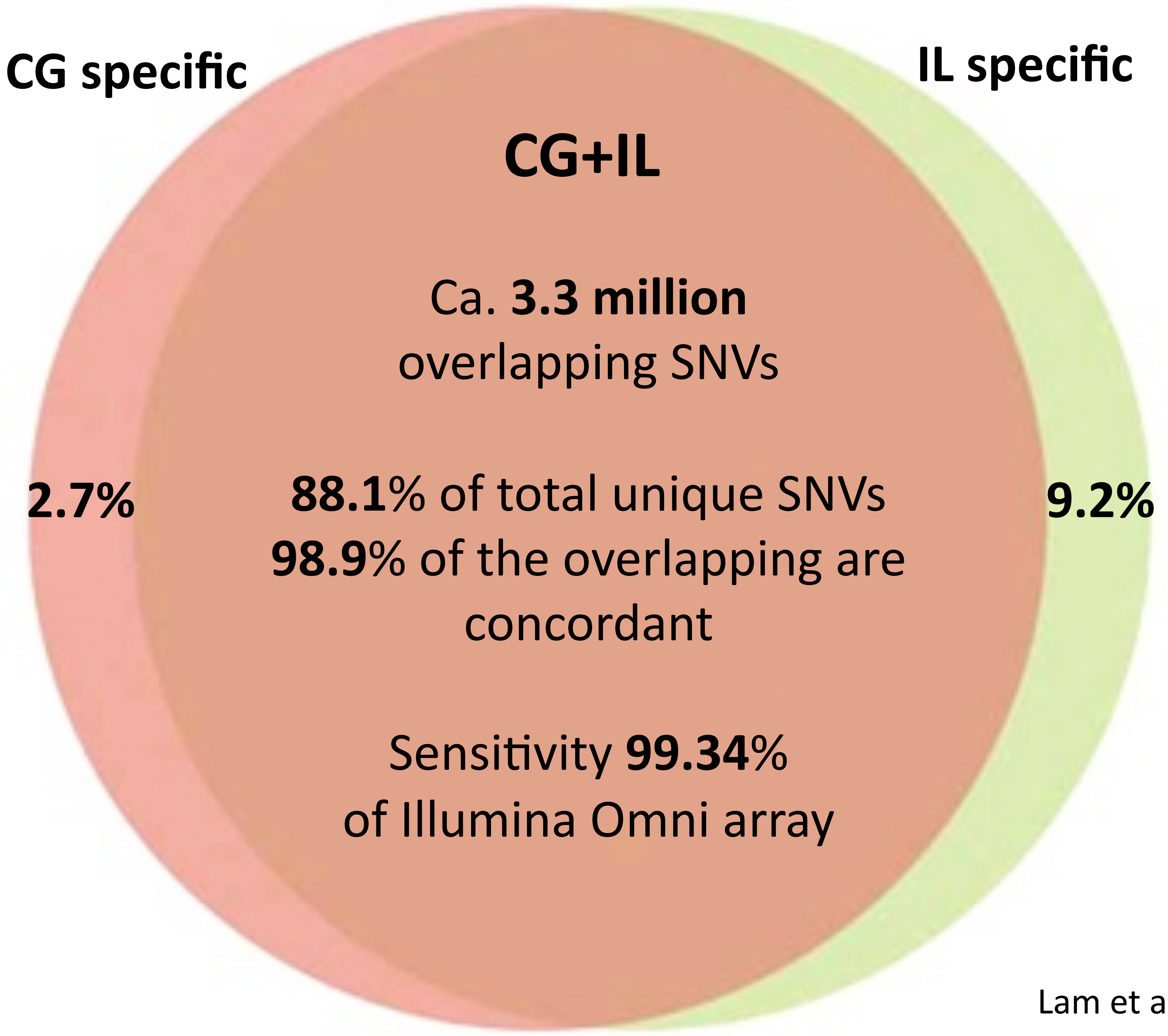
Sanger sequencing: randomly selected SNVs, both concordant and platform specific

- Validated 20 of 20 concordant SNVs; 2 of 15 (13.3%) Illumina-specific and 17 of 18 (94.4%) CG-specific

Agilent SureSelect target enrichment for total 33,084 SNVs

- Validated 92.7% of concordant; 61.9% CG-specific; 64.3% Illumina-specific SNVs

Lam et al. 2012



CG specific

IL specific

CG+IL

Ca. **3.3 million** overlapping SNVs

2.7%

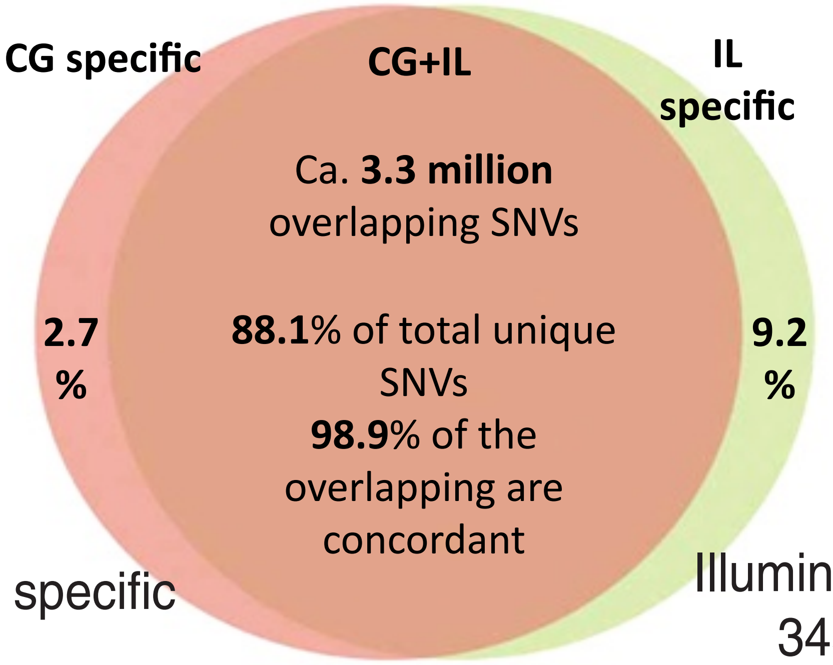
88.1% of total unique SNVs
98.9% of the overlapping are concordant

9.2%

Sensitivity **99.34%** of Illumina Omni array

Lam et al. 2012

Lam et al. 2012



CG specific

CG+IL

IL specific

Ca. **3.3 million** overlapping SNVs

88.1% of total unique SNVs

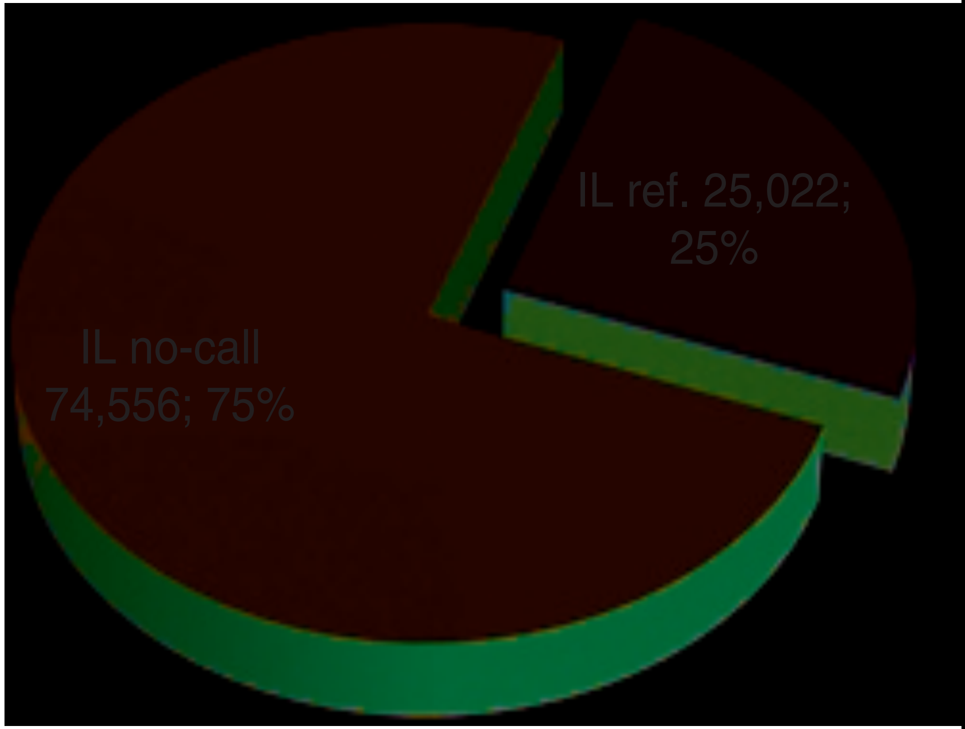
98.9% of the overlapping are concordant

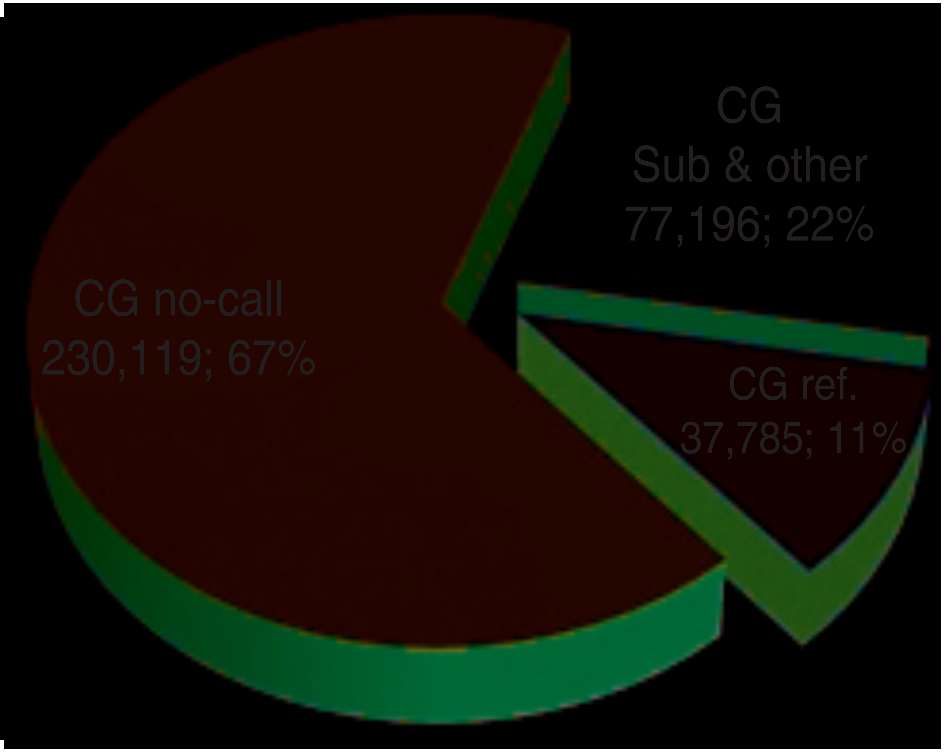
2.7 %

9.2 %

Complete Genomics specific 99,578

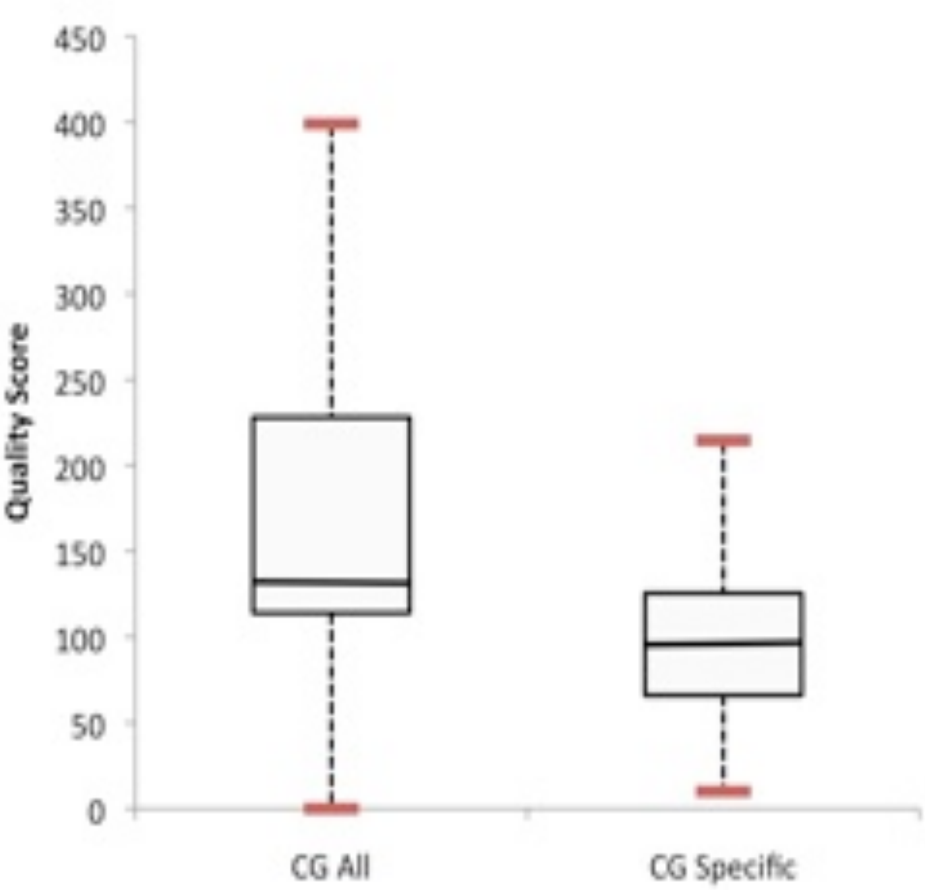
Illumina specific 345,100

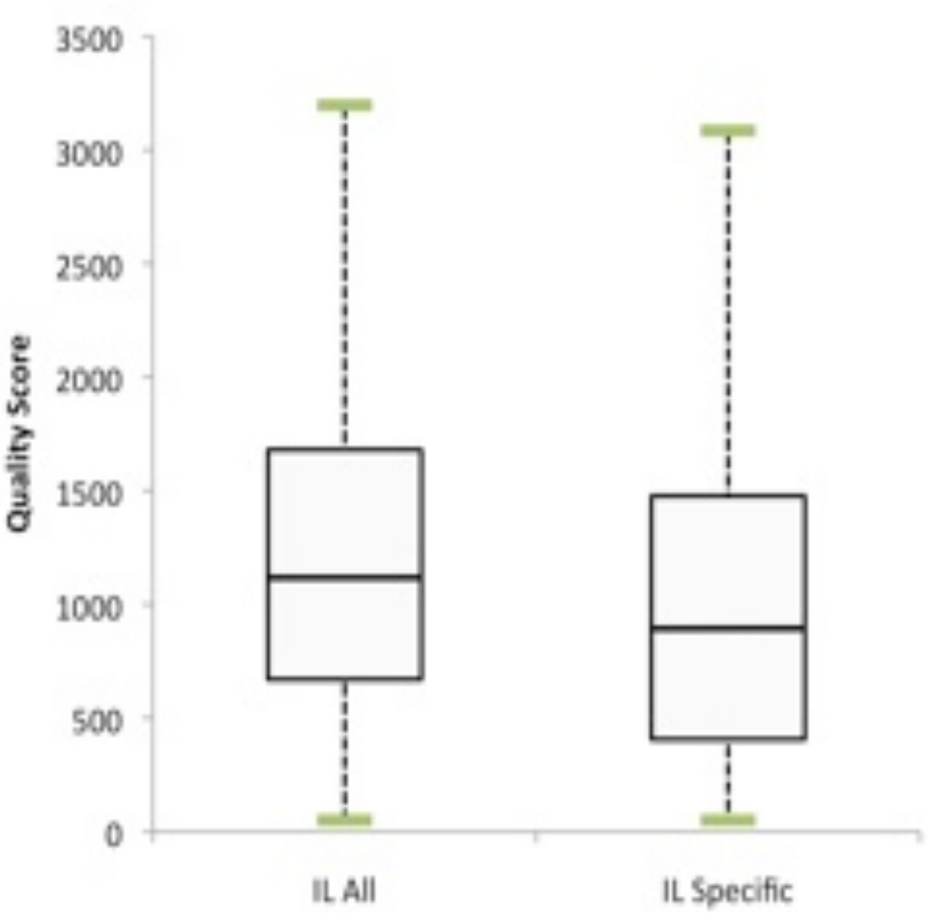




- ti/tv of concordant SNVs was very close to that expected, whereas the platform-specific ti/tv was much lower

- Quality scores of platform-specific SNVs were lower





Lam et al. 2012

- Average GC content
    - Concordant: 0.46
    - CG-specific: 0.45
    - Illumina-specific: 0.41

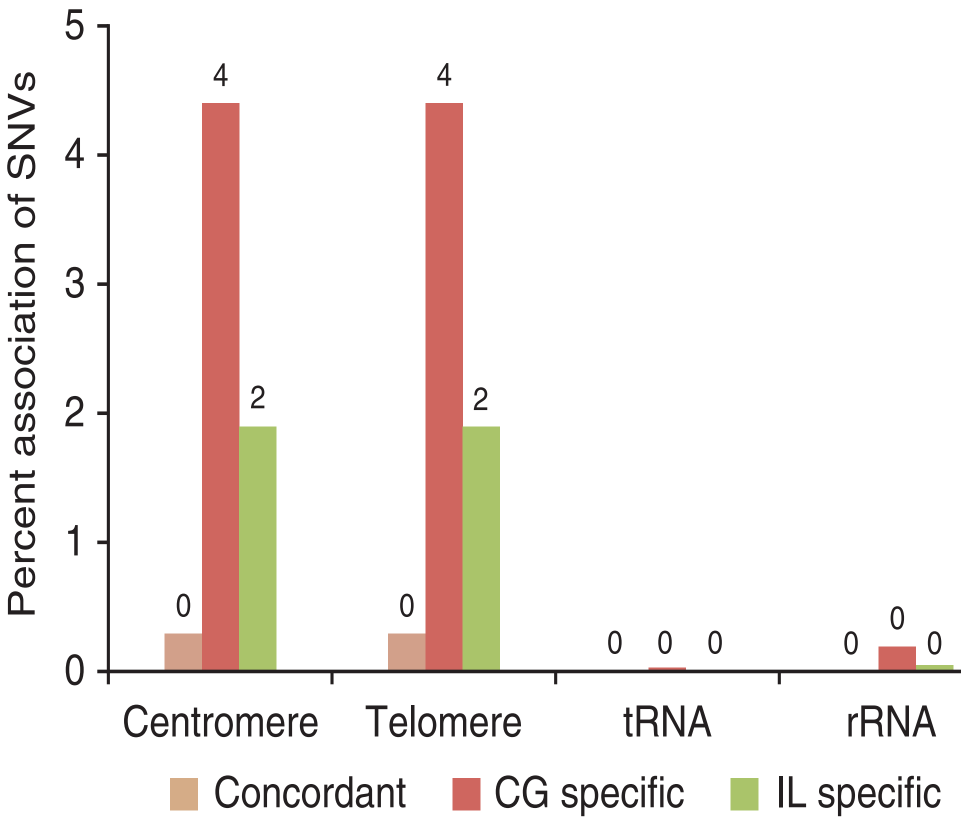
- Average read depths: 48, 47 and 44, respectively.

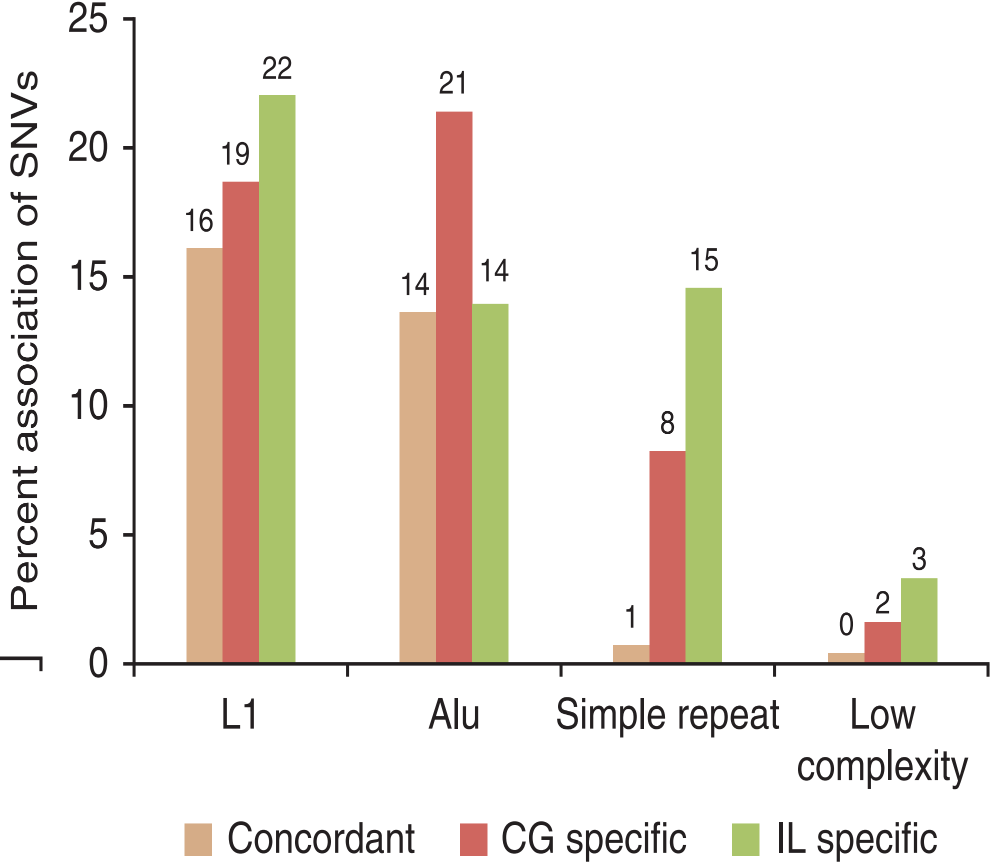
- No strong correlation of SNV detection with GC content.

Lam et al. 2012

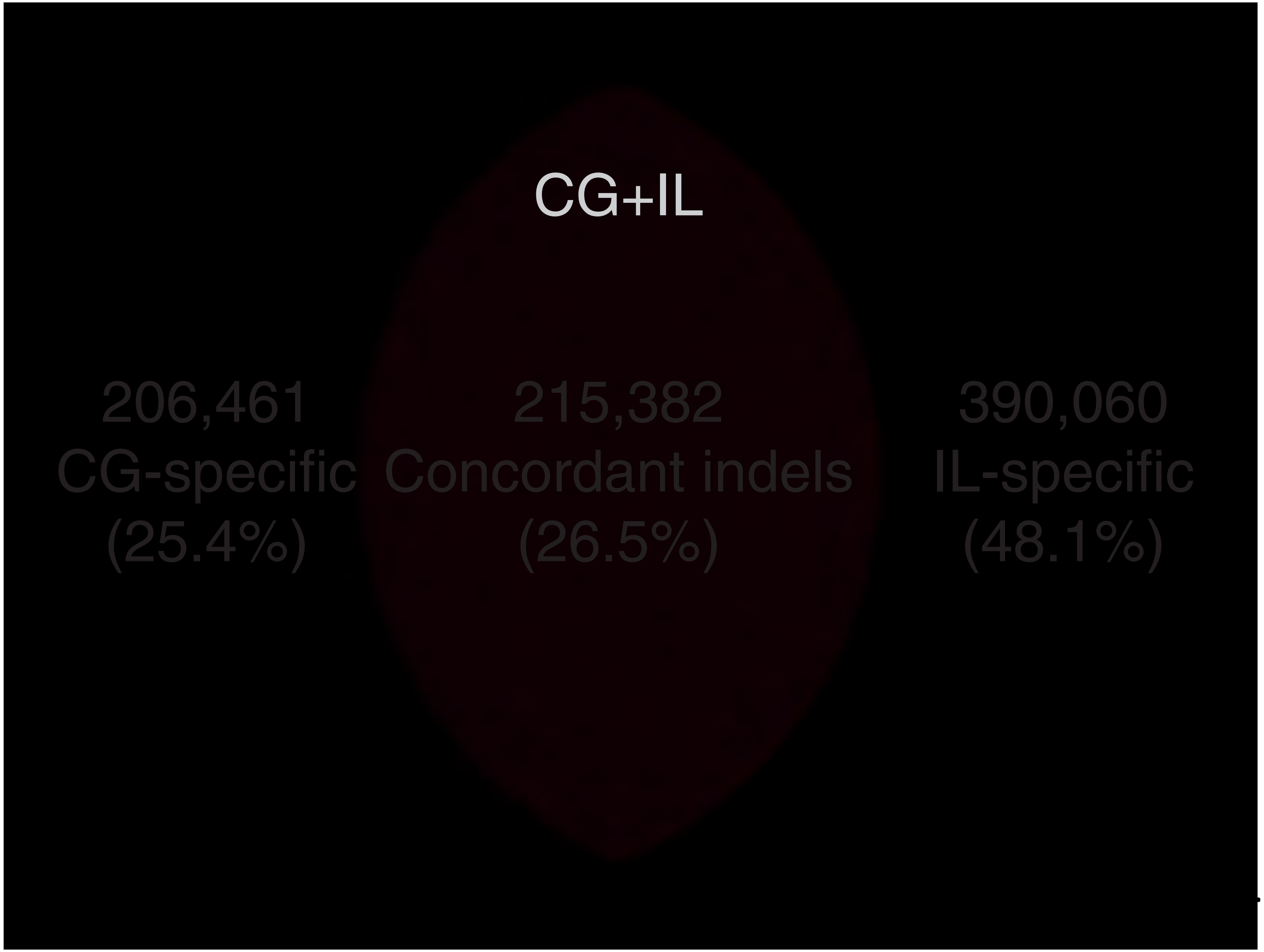
- Platform-specific SNVs associated with repetitive elements such as Alu, telomere and simple repeat sequences
- Only 0.3% of the concordant SNVs in telomeres or centromeres, opposed to 4% of CG-specific and 2% of Illumina-specific

Lam et al. 2012





Indel calls are greatly discordant



CG+IL

206,461
CG-specific
(25.4%)

215,382
Concordant indels
(26.5%)

390,060
IL-specific
(48.1%)

. 2012

Conclusions

- Each genome sequencing approach is generally capable of detecting most SNVs
- CG appears to be more accurate, but also slightly less sensitive
- llumina covers more bases and makes a higher number of overall calls, but also has more false positives

Lam et al. 2012

Conclusions

- Each genome sequencing approach is generally capable of detecting most SNVs
- CG appears to be more accurate, but also slightly less sensitive
- llumina covers more bases and makes a higher number of overall calls, but also has more false positives
- Both methods clearly call variants missed by the other technology
- Filtering for repeat regions and quality helps to reduce errors
- For maximized sensitivity and quality, sequence in parallel

Lam et al. 2012

OPEN ACCESS Freely available online

PLOS | ONE

Comparison of Sequencing Platforms for Single Nucleotide Variant Calls in a Human Sample

Aakrosh Ratan[1¶], Webb Miller[1], Joseph Guillory[2], Jeremy Stinson[2], Somasekar Seshagiri[2], Stephan C. Schuster[1*¶]

1 Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Department of Molecular Biology, Genentech Inc., South San Francisco, California, United States of America

- 454/Roche GS FLX
- Illumina HiSeq 2000
- ABI SOLiD 3 ECC

- Overlap of SNVs

- 1 DNA sample

Ratan et al. 2013

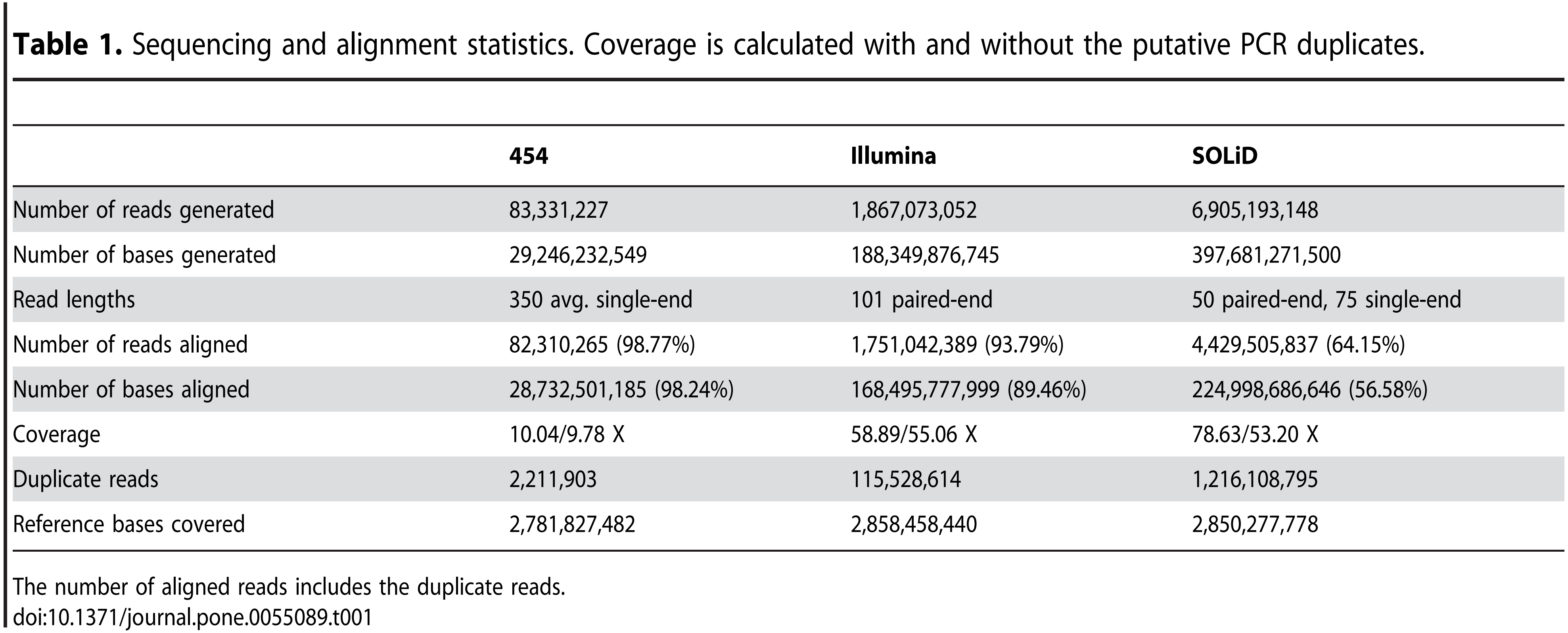


Table 1. Sequencing and alignment statistics. Coverage is calculated with and without the putative PCR duplicates.

	454	**Illumina**	**SOLiD**
Number of reads generated	83,331,227	1,867,073,052	6,905,193,148
Number of bases generated	29,246,232,549	188,349,876,745	397,681,271,500
Read lengths	350 avg. single-end	101 paired-end	50 paired-end, 75 single-end
Number of reads aligned	82,310,265 (98.77%)	1,751,042,389 (93.79%)	4,429,505,837 (64.15%)
Number of bases aligned	28,732,501,185 (98.24%)	168,495,777,999 (89.46%)	224,998,686,646 (56.58%)
Coverage	10.04/9.78 X	58.89/55.06 X	78.63/53.20 X
Duplicate reads	2,211,903	115,528,614	1,216,108,795
Reference bases covered	2,781,827,482	2,858,458,440	2,850,277,778

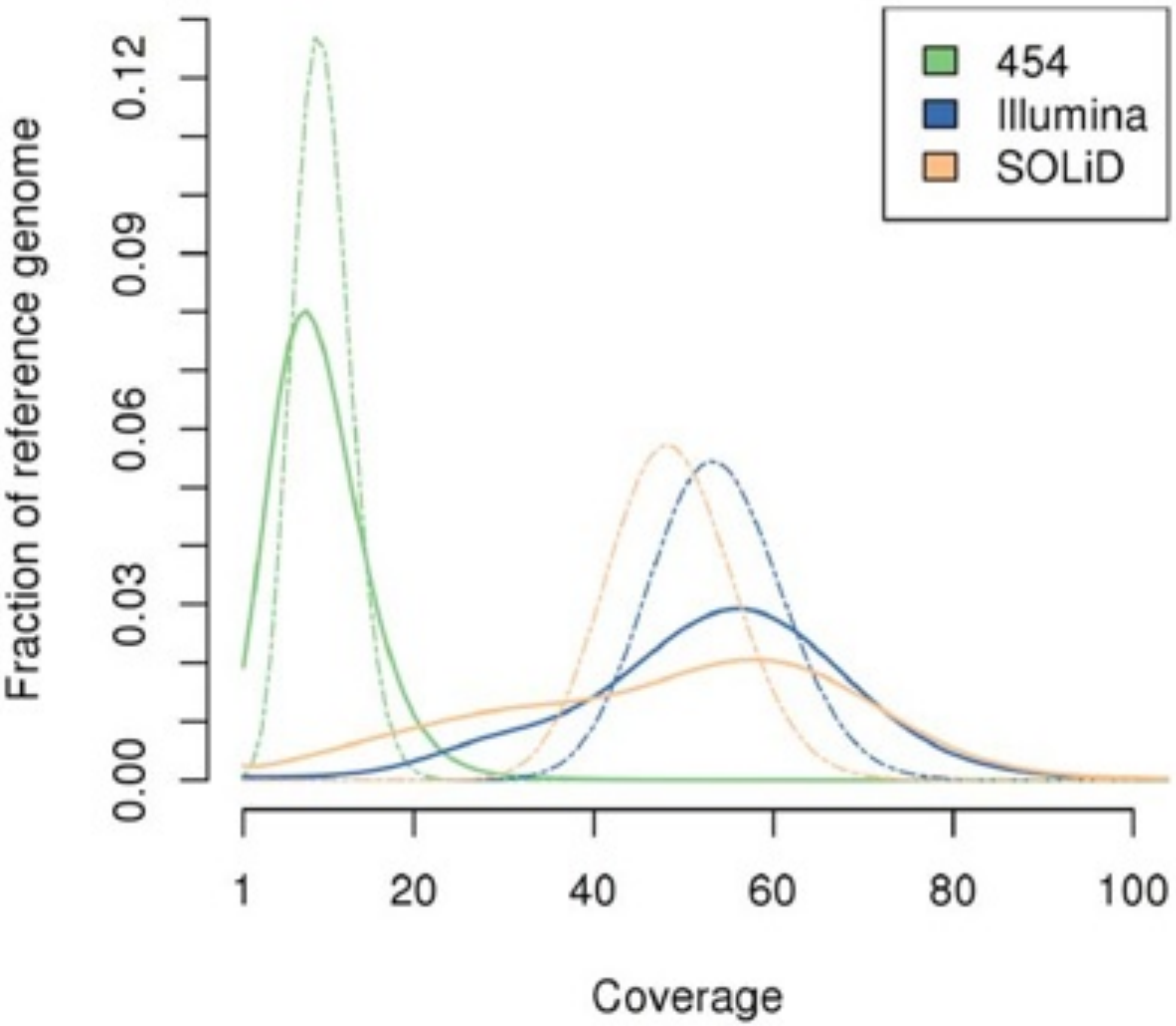
The number of aligned reads includes the duplicate reads.
doi:10.1371/journal.pone.0055089.t001

- Uneven coverage between platforms
- “we sequenced the individual’s DNA to read-depths that allows for variant detection in each corresponding dataset with sufficient confidence “

Ratan et al. 2013

Coverage

- 454: 10x
- Illumina: 59x
- SOLiD: 79x



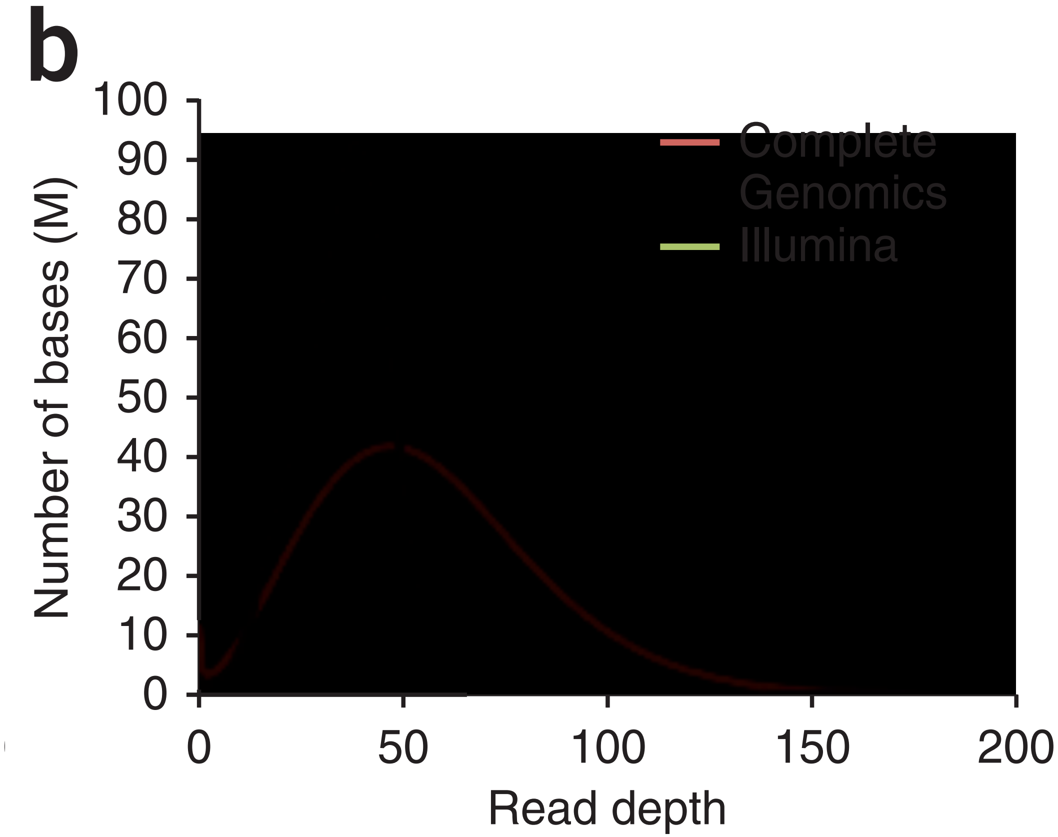
Removing the sex chromosomes from the analysis did not eliminate the bimodal behavior

Ratan et al. 2013

Lam et al. 2012:

Average coverage 76x

CG less uniform in coverage



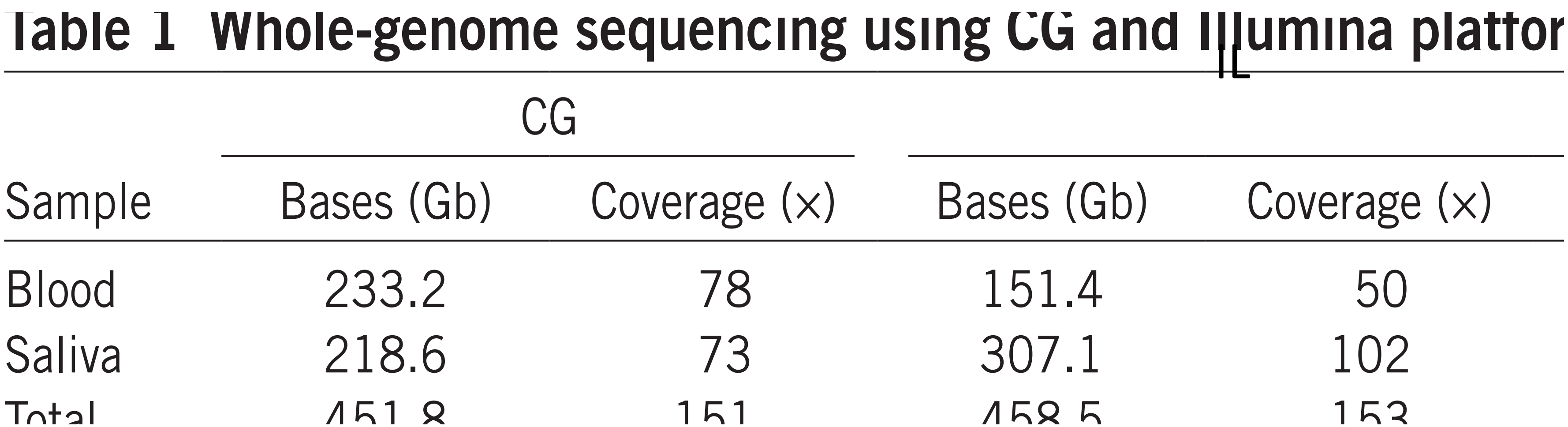
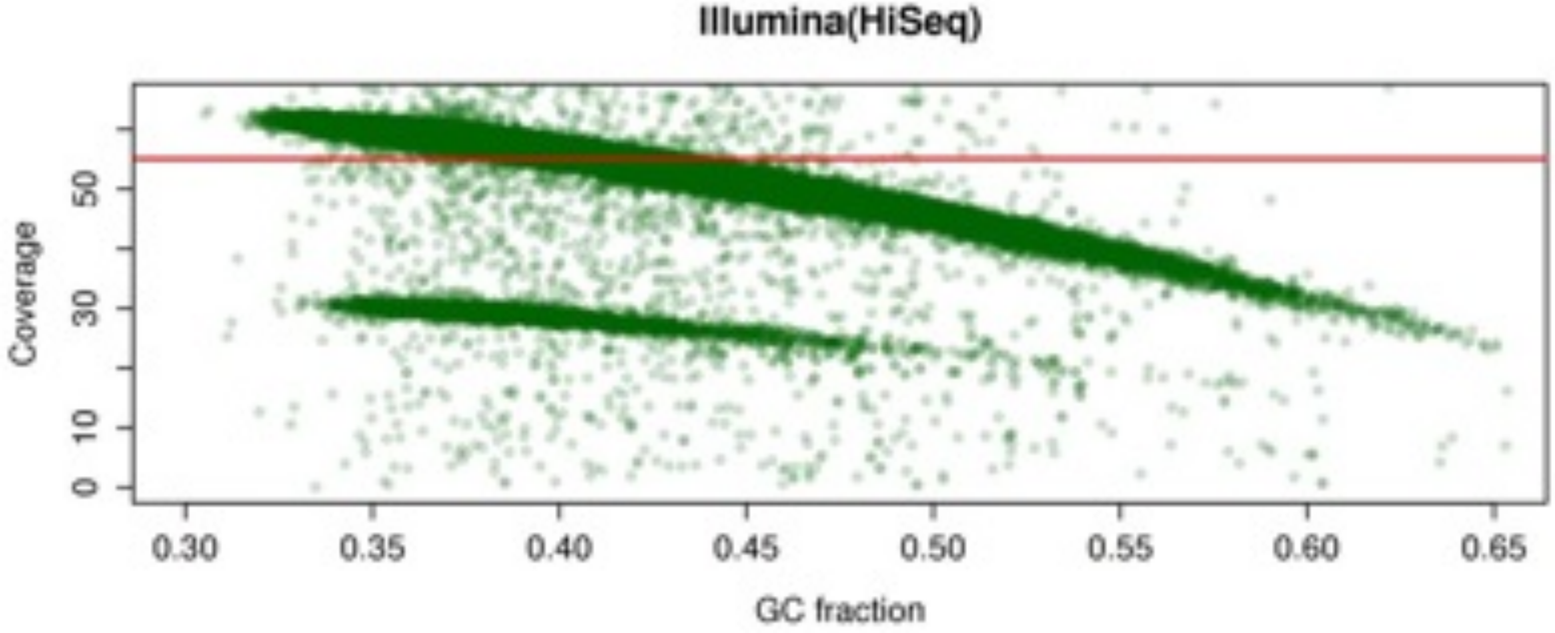


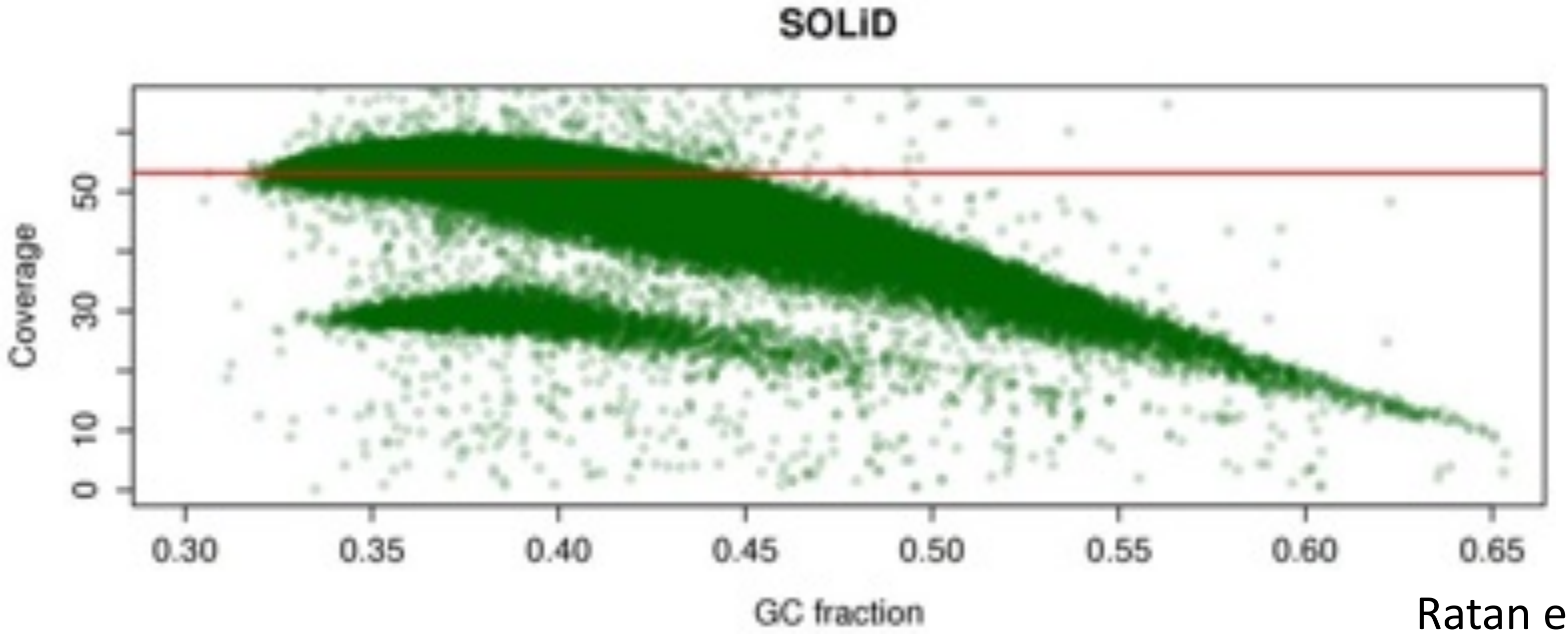
Table 1 Whole-genome sequencing using CG and Illumina platfor

	CG		IL	
Sample	Bases (Gb)	Coverage (×)	Bases (Gb)	Coverage (×)
Blood	233.2	78	151.4	50
Saliva	218.6	73	307.1	102
Total	451.8	151	458.5	153

Lam et al. 2012

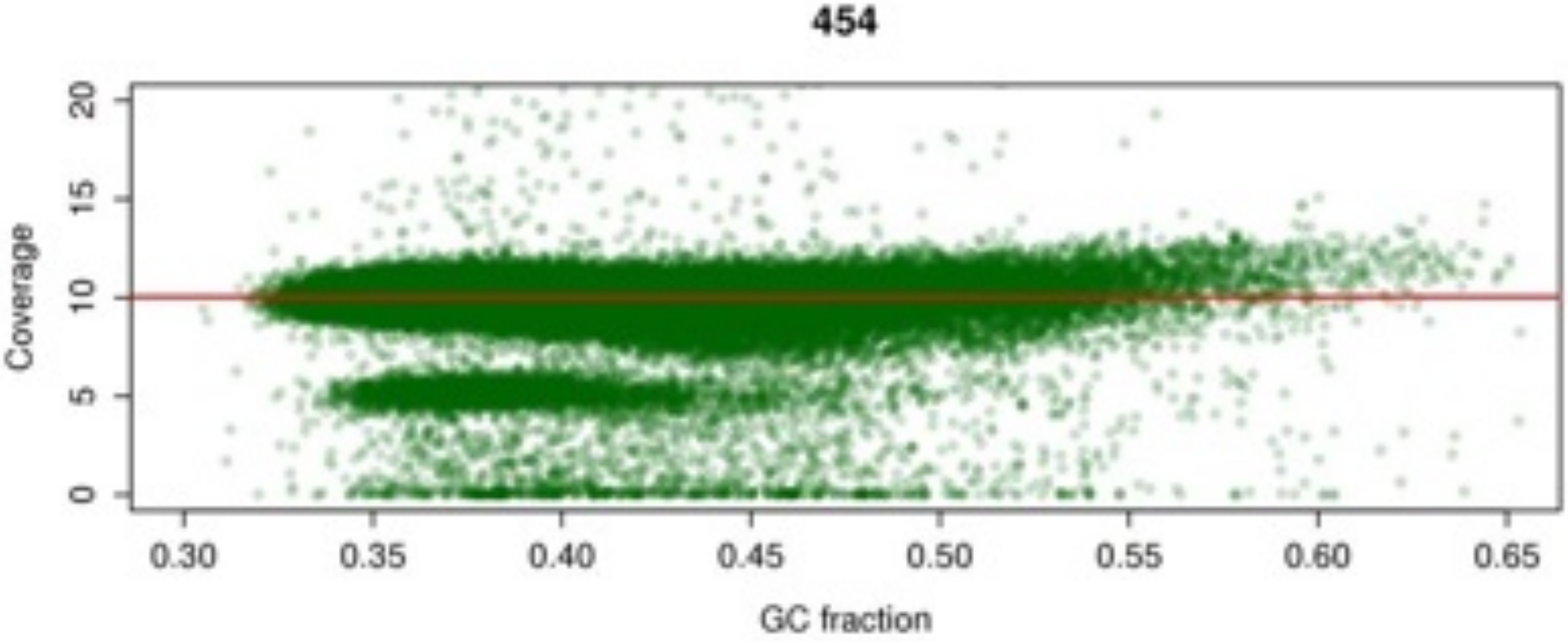
Variation of coverage with GC content





Ratan et al. 2013

Variation of coverage with GC content



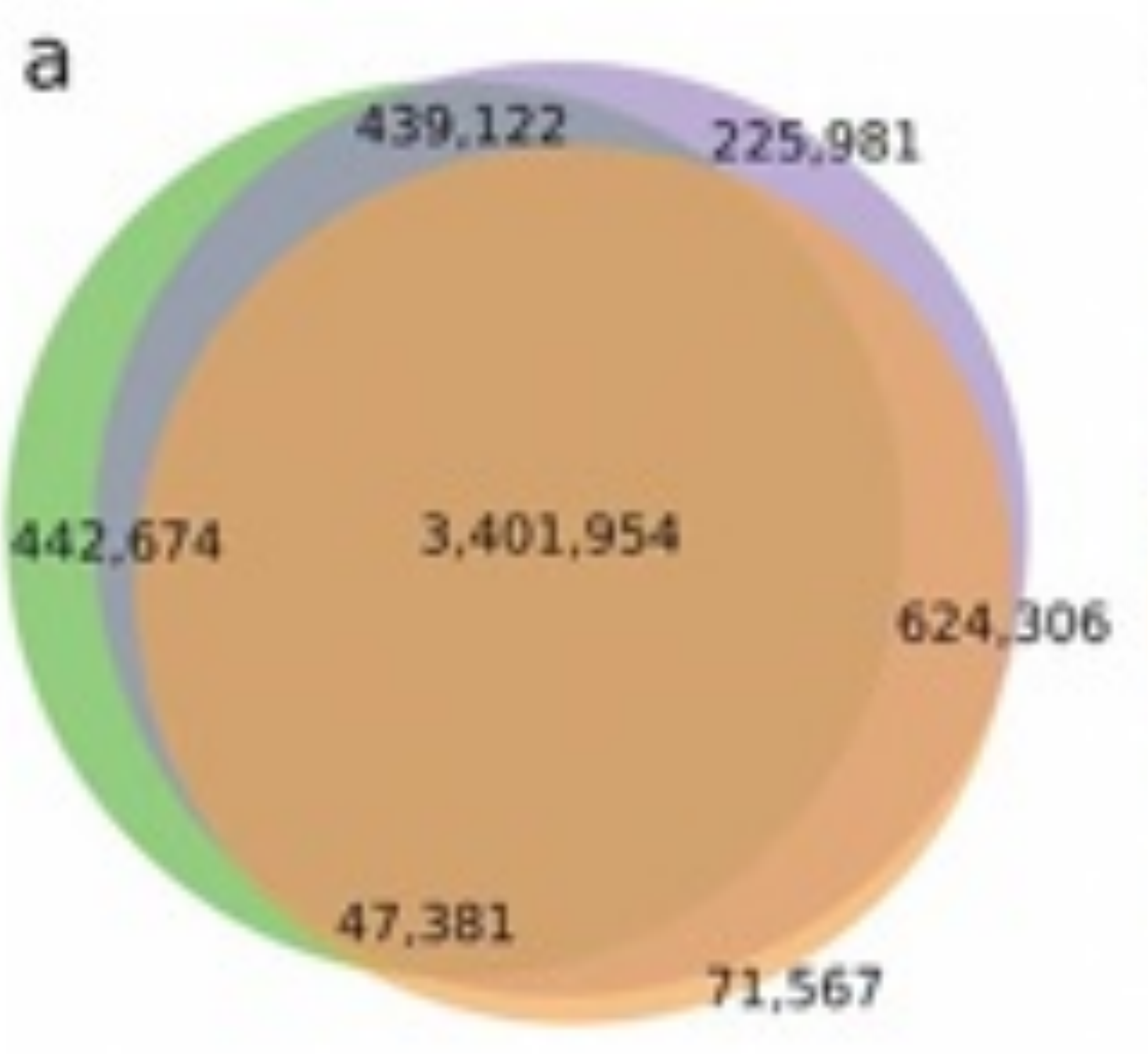
Ratan et al. 2013

Numbers of detected variants

- 454: 4,331,131
- Illumina: 4,691,363
- SOLiD: 4,145,208
- Total combined: 5,252,985
- Shared between three: 3,401,954 (64.8%)

Ratan et al. 2013

Numbers of detected variants – 5.25 M total

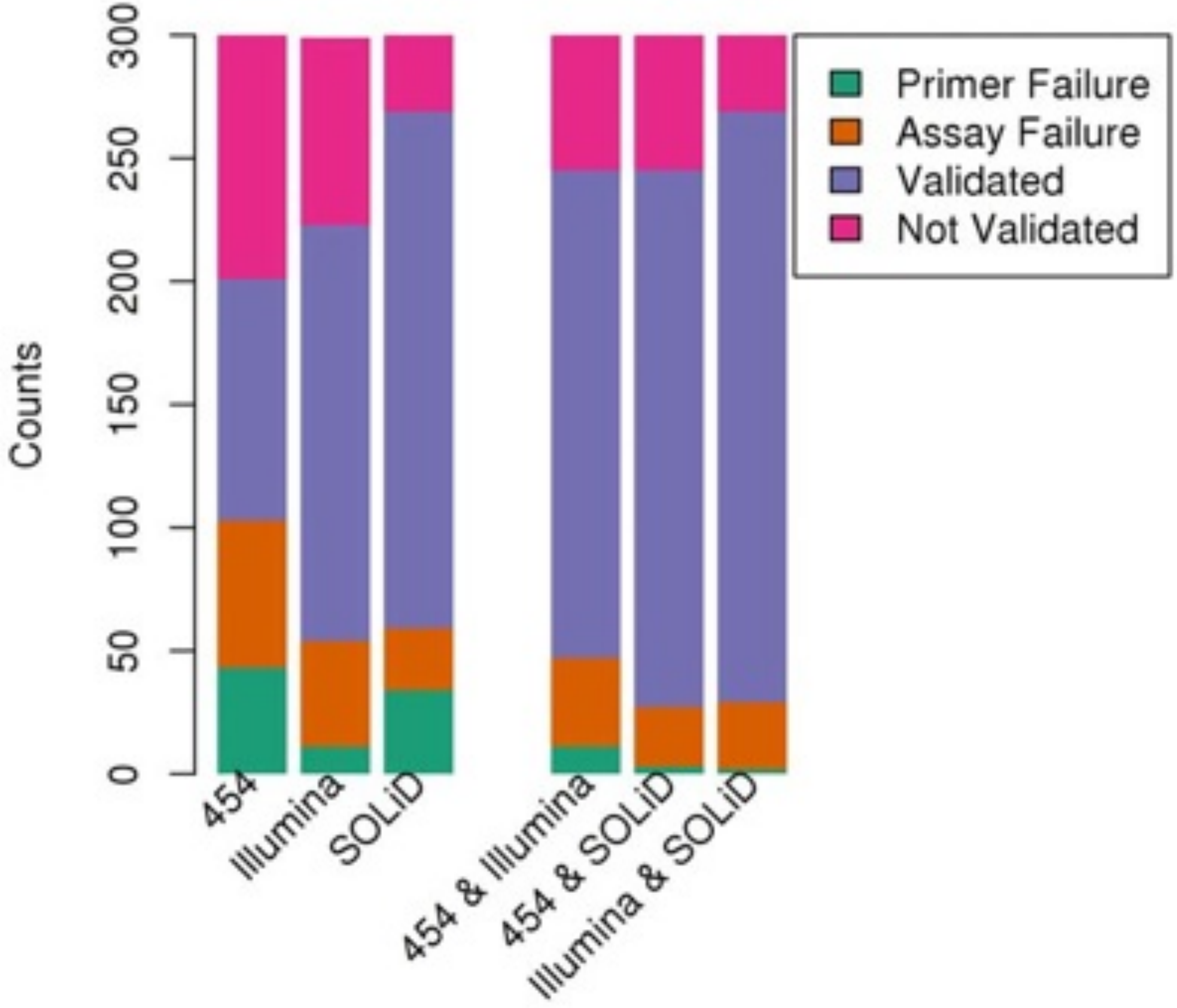


Ratan et al. 2013

- “self-masking” to identify the regions in the reference genome where reads should align uniquely – LASTZ
- reference genome broken into fragments of length equal to the length of the reads
- Fragments mapped back to reference
- Considering only uniquely mappable regions
- Very useful for 454 data
- Less so for Illumina and SOLiD

Ratan et al. 2013

Validation using Sequenom Mass Spectrometry



Ratan et al. 2013

Reducing platform bias in next-generation sequencing.

Lauri Saag[1,2], Ulvi Gerst Talas[1], Mario Mitt[1,3], Reedik Mägi[3], Simon Rasmussen[4], Richard Villems[1,2] and Mait Metspalu[2]

Poster at ASHG 2013 in Boston

Complete Genomics vs. Illumina

Concordance of base calls in potential SNP positions

Study design.

- Genomes of 7 individuals sequenced on both platforms.
- Concordance measurements presented for 4 samples with good and similar sequence statistics.
- CG: average coverage 40x, processed with CG pipeline.
- Illumina: average coverage 26x, mapping with BWA, multisample calling with SAMtools.

Study design.

- Positions selected where SNP occurred in at least one individual in a dataset of 249 humans from various populations worldwide (sequenced by CG). Sites with other types of mutations excluded.

- From this list of ca. 50 million SNP sites, we selected those for which information (either reference, variant or no-call) was available from both platforms and all samples.

- **Thus, our comparison is based on the call concordance in predefined sites, and not on the overlap of variant-only data.** If a site is discordant between platforms, it appears as private for both.

Initial filtering.

- CG: VQHIGH (PASS) filter.
- Illumina: Read depth min 10, max 100; males X, Y chromosomes min 5, max 50. **Variants in repeats removed.**
- Ambiguous bases (N) in compared files may be either no-calls or filtered.

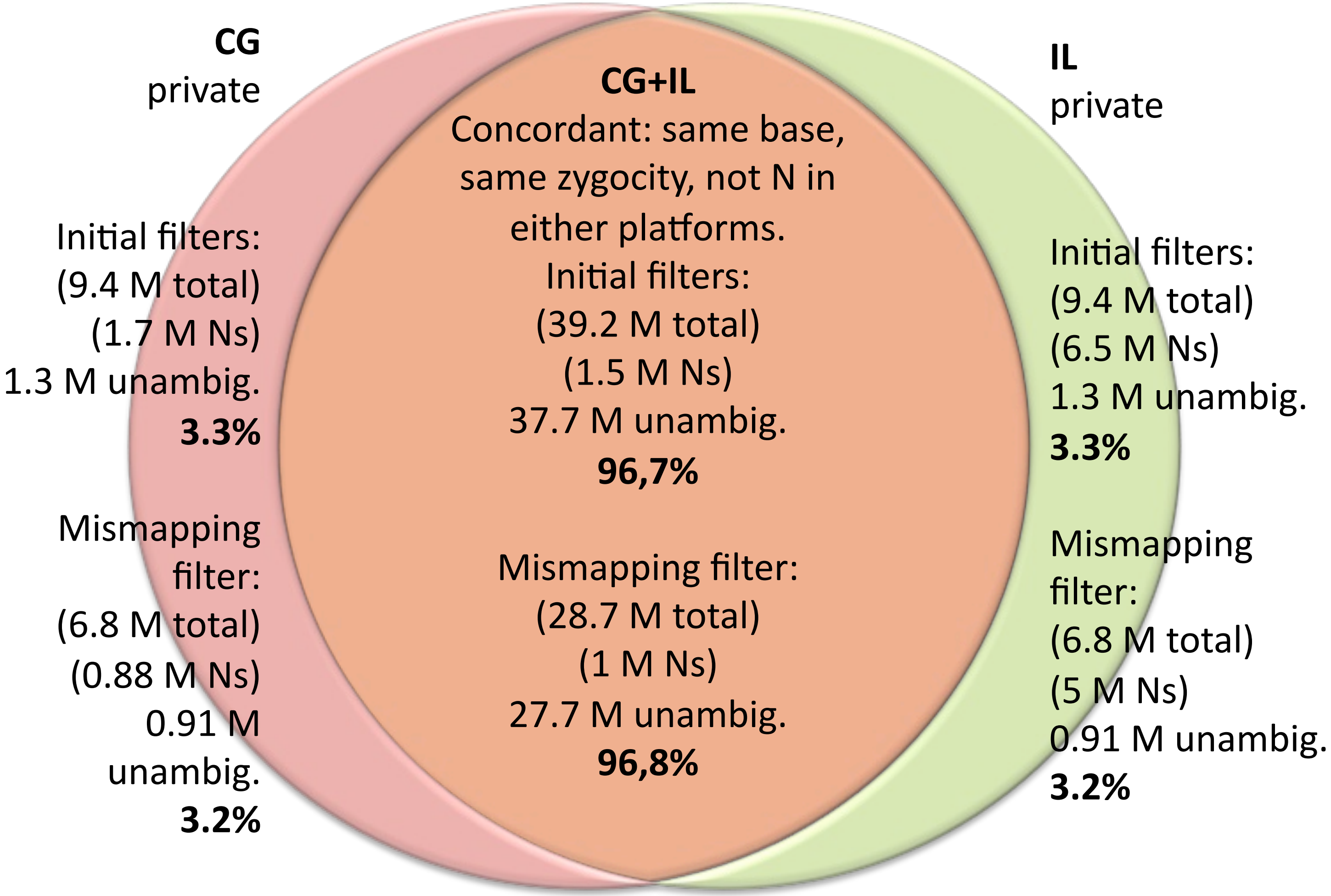
Initial filtering.

- CG: VQHIGH (PASS) filter.
- Illumina: Read depth min 10, max 100; males X, Y chromosomes min 5, max 50. **Variants in repeats removed.**
- Ambiguous bases (N) in compared files may be either no-calls or filtered.

Mismapping positions filter.

- We masked the data with windows in the genome where mapping of Illumina reads often fails.
- Applicability of these blacklists to CG data was not known.

- Total 48.6 M sites, 39 M unambiguous (not N) in both platforms.
- 29.2 M sites concordant in all 4 samples, 28.8 M unambiguous.
- **Averages over 4 samples**, given for unambiguous discordances (percent ranges ±0.1)



CG
private

Initial filters:
(9.4 M total)
(1.7 M Ns)
1.3 M unambig.
3.3%

Mismapping filter:
(6.8 M total)
(0.88 M Ns)
0.91 M unambig.
3.2%

CG+IL
Concordant: same base, same zygocity, not N in either platforms.

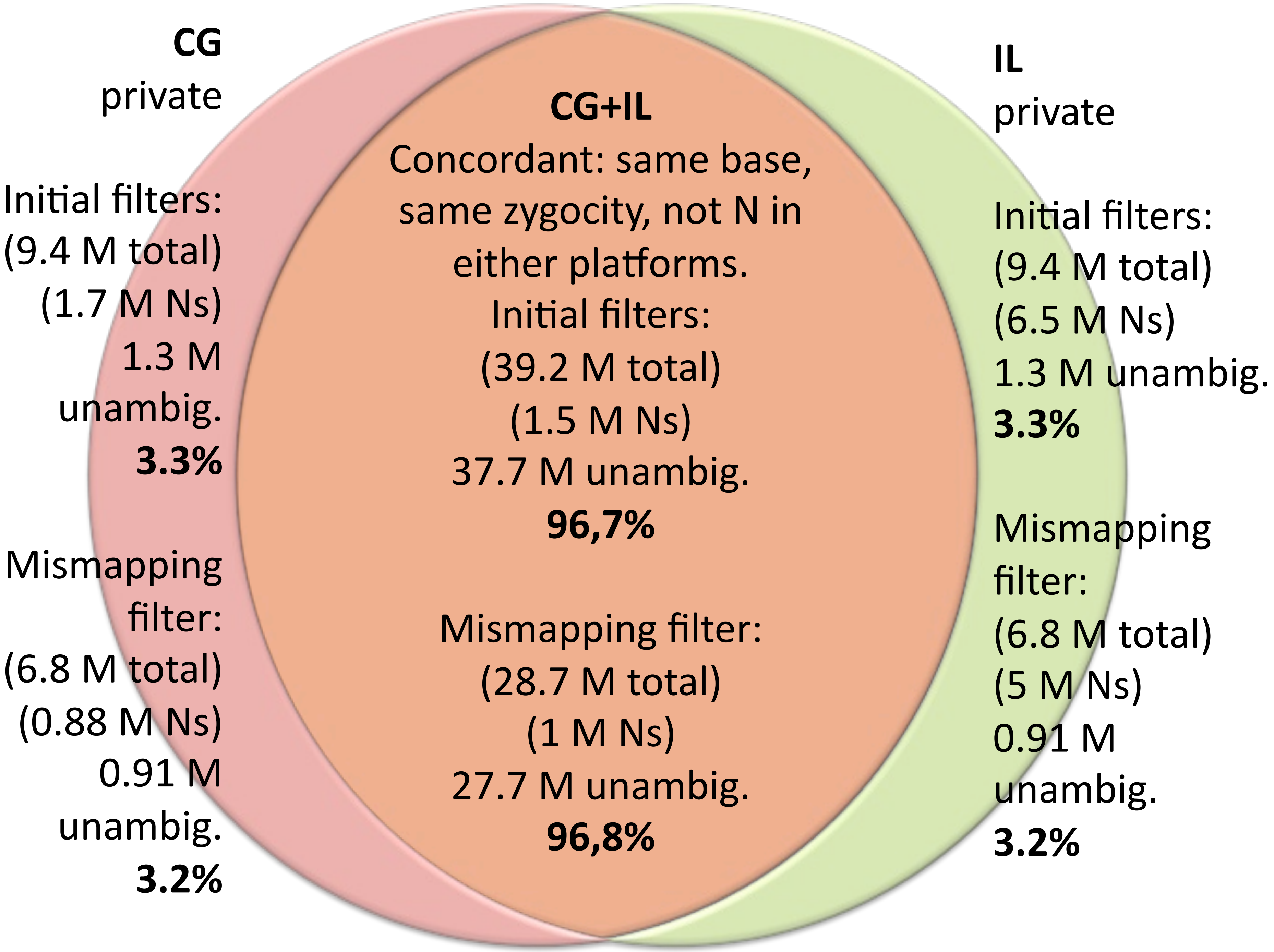
Initial filters:
(39.2 M total)
(1.5 M Ns)
37.7 M unambig.
96,7%

Mismapping filter:
(28.7 M total)
(1 M Ns)
27.7 M unambig.
96,8%

IL
private

Initial filters:
(9.4 M total)
(6.5 M Ns)
1.3 M unambig.
3.3%

Mismapping filter:
(6.8 M total)
(5 M Ns)
0.91 M unambig.
3.2%


CG
private
Initial filters:
(9.4 M total)
(1.7 M Ns)
1.3 M
unambig.
3.3%
Mismapping
filter:
(6.8 M total)
(0.88 M Ns)
0.91 M
unambig.
3.2%
CG+IL
Concordant: same base,
same zygocity, not N in
either platforms.
Initial filters:
(39.2 M total)
(1.5 M Ns)
37.7 M unambig.
96,7%
Mismapping filter:
(28.7 M total)
(1 M Ns)
27.7 M unambig.
96,8%
IL
private
Initial filters:
(9.4 M total)
(6.5 M Ns)
1.3 M unambig.
3.3%
Mismapping
filter:
(6.8 M total)
(5 M Ns)
0.91 M
unambig.
3.2%

- **Poor-mapping windows only minimally reduce the bias.**
- The bias in these positions is not greater than elsewhere, suggesting that if errors are indeed present in these positions in our Illumina data then the current version of our windows may also apply to CG.

Concordant positions filter and additional 3 samples.

- 29.2 M sites were concordant in all 4 samples.
- We used these as a whitelist to filter the sequences of other 3 samples that had lower average quality and large numbers of ambiguous sites.

Concordant positions filter and additional 3 samples.

- 29.2 M sites were concordant in all 4 samples.
- We used these as a whitelist to filter the sequences of other 3 samples that had lower average quality and large numbers of ambiguous sites.
- **On average, from 22.3 M unambiguous sites in both platforms 18.1 remained (19%) and the bias dropped to 1.4% from 3.4%.**
- **Despite lower quality, the additional 3 samples showed stable and very similar bias percentage as the other four:** 3.4% after initial filtering, 3.3% with mismapping filter and 3.1% with biased positions filter.