Perfecting the human genome with longer reads

Mauricio Carneiro, Sheila Fisher and Mark DePristo

Group Lead - TechDev
Genome Sequencing and Analysis Group
Program in Medical and Population Genetics
Broad Institute of Harvard and MIT

What are the challenges we face today?

- a large amount of the genome remains consistently uncovered with the current technology
- we spend a lot of our sequencing budget to bring poorly covered regions to minimum analyzable coverage. (especially in whole exome projects)
- these are due to bias in sequencing, capture and alignment bias (see Michael Ross poster on friday)

What can reduce bias?

- Different library prep techniques can significantly help (see Mark DePristo's talk on friday)
- Different capture approaches
- Longer sequencing reads

Low covered sites are a cost burden

- sites that are consistently ***low covered*** force the sequencing center to over-sequence
  - For targeted sequencing (e.g. WEx) this also means bait balancing and other costly procedures.
  - currently at least 14% of the genome is consistently low covered

Uncovered sites are a true analytical challenge

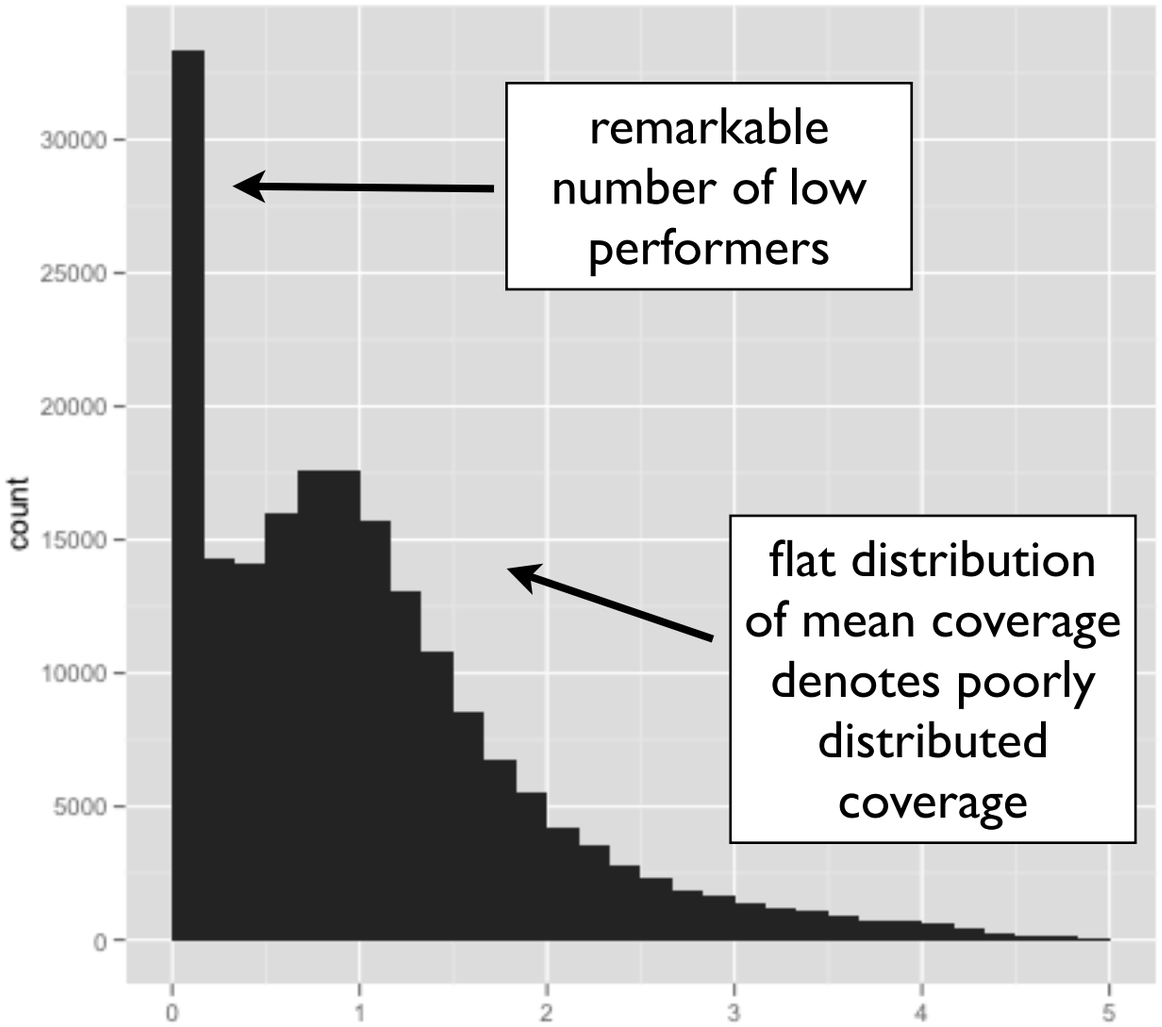
- sites that are consistently ***uncovered*** cannot be analyzed (who knows what's in there?)
- there is no alternative solution (we can't throw money to solve this)
- currently at least 6.1% of the genome is consistently uncovered

Data and definitions

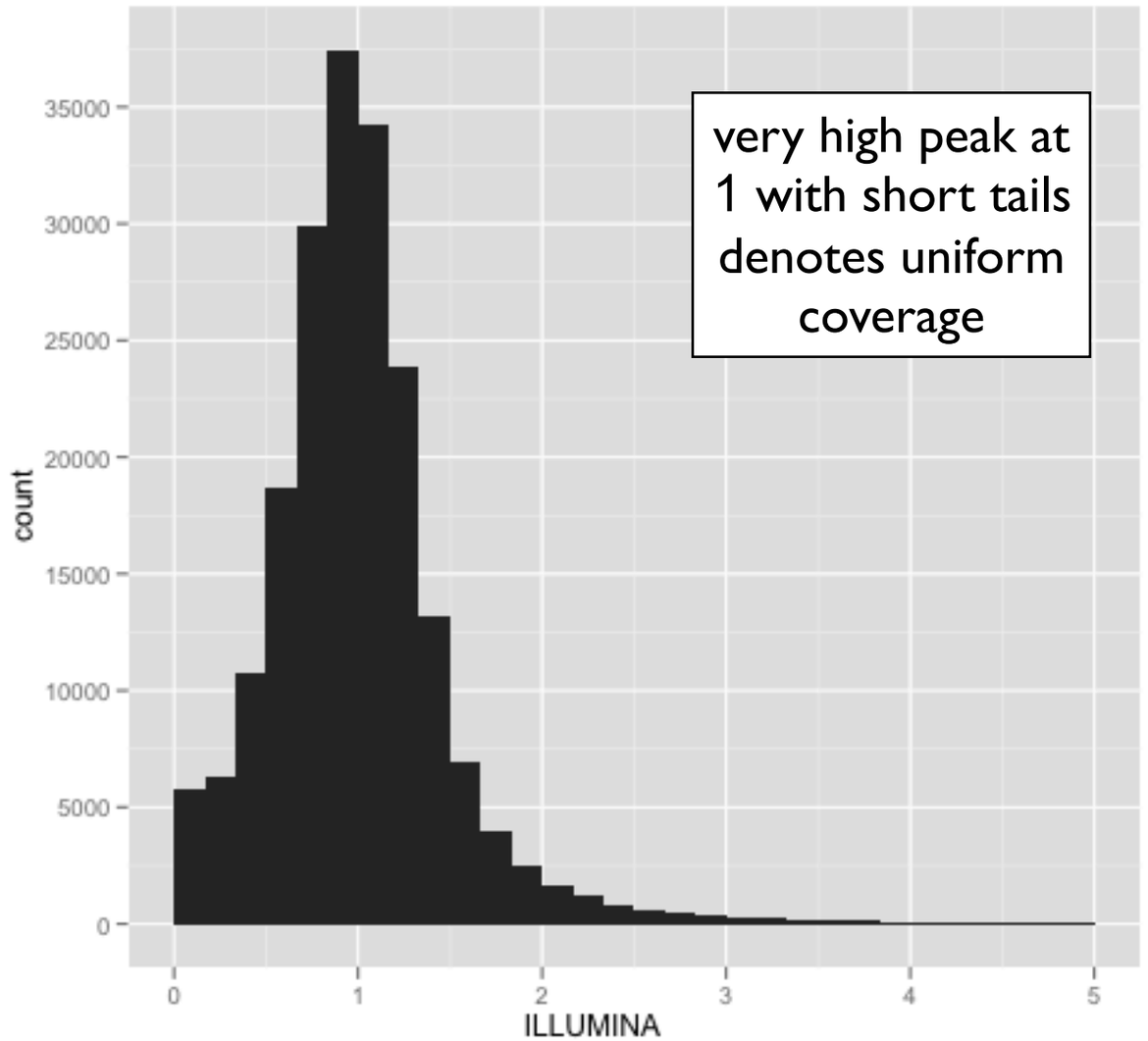
- For target bias we used a comparison between Illumina's Nextera Rapid Capture Exome product against our current production protocol
- For sequencing and alignment bias we analyzed PCRFree datasets of different read lengths (2x250, 2x101 and 1x32) produced at the Broad on Illumina HiSeq 2500.
- Variant calling sensitivity was measured on the same sequence data for NA12878 and compared against the NA12878 knowledge base truth dataset (Broad internal validation dataset)

Reducing bias in whole exome (targeted) projects

Illumina's Nextera Rapid Capture Exome product significantly reduced target bias

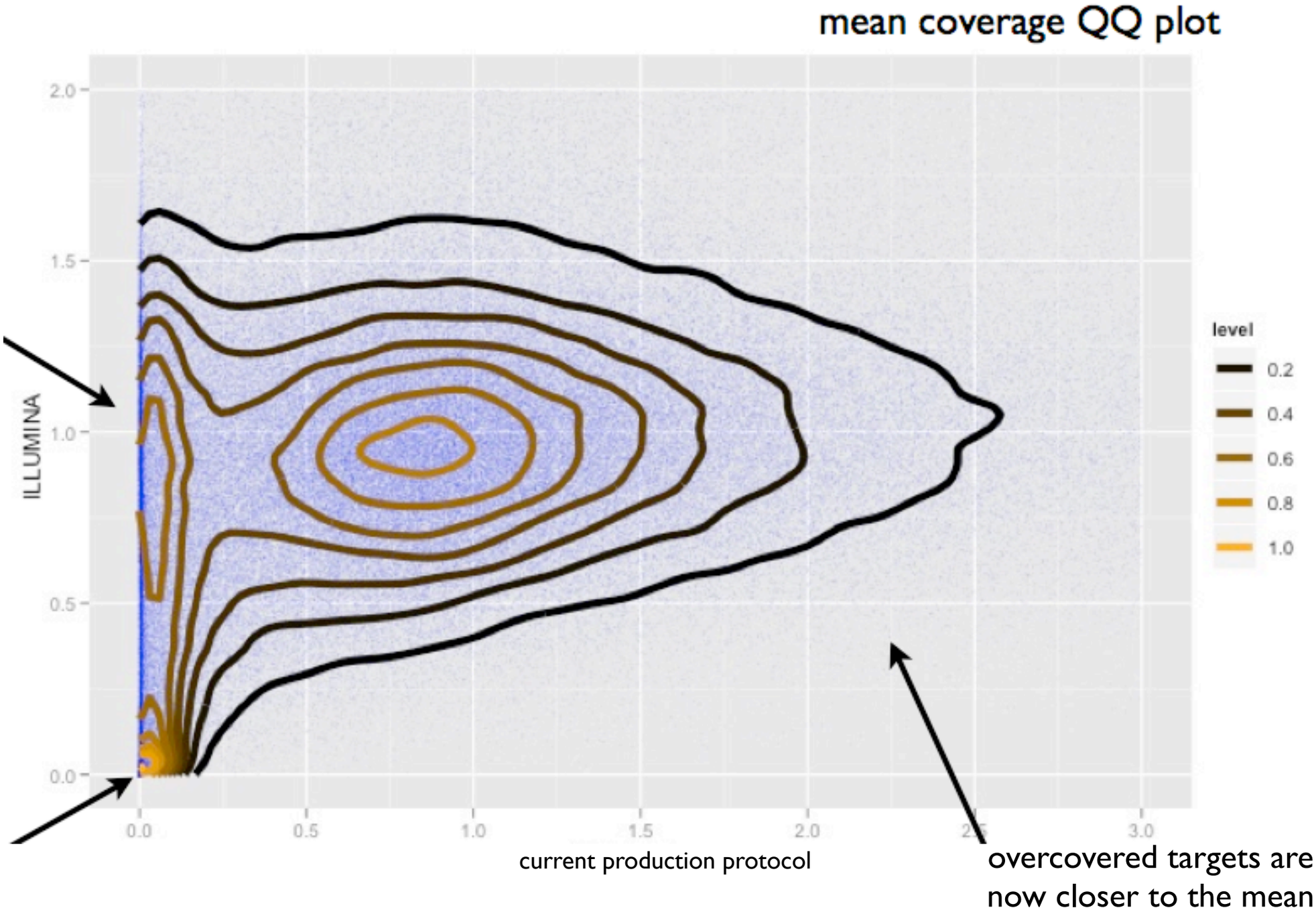


current production protocol



Nextera Rapid Capture Exome

Illumina's Nextera Rapid Capture Exome product covers targets we couldn't cover before

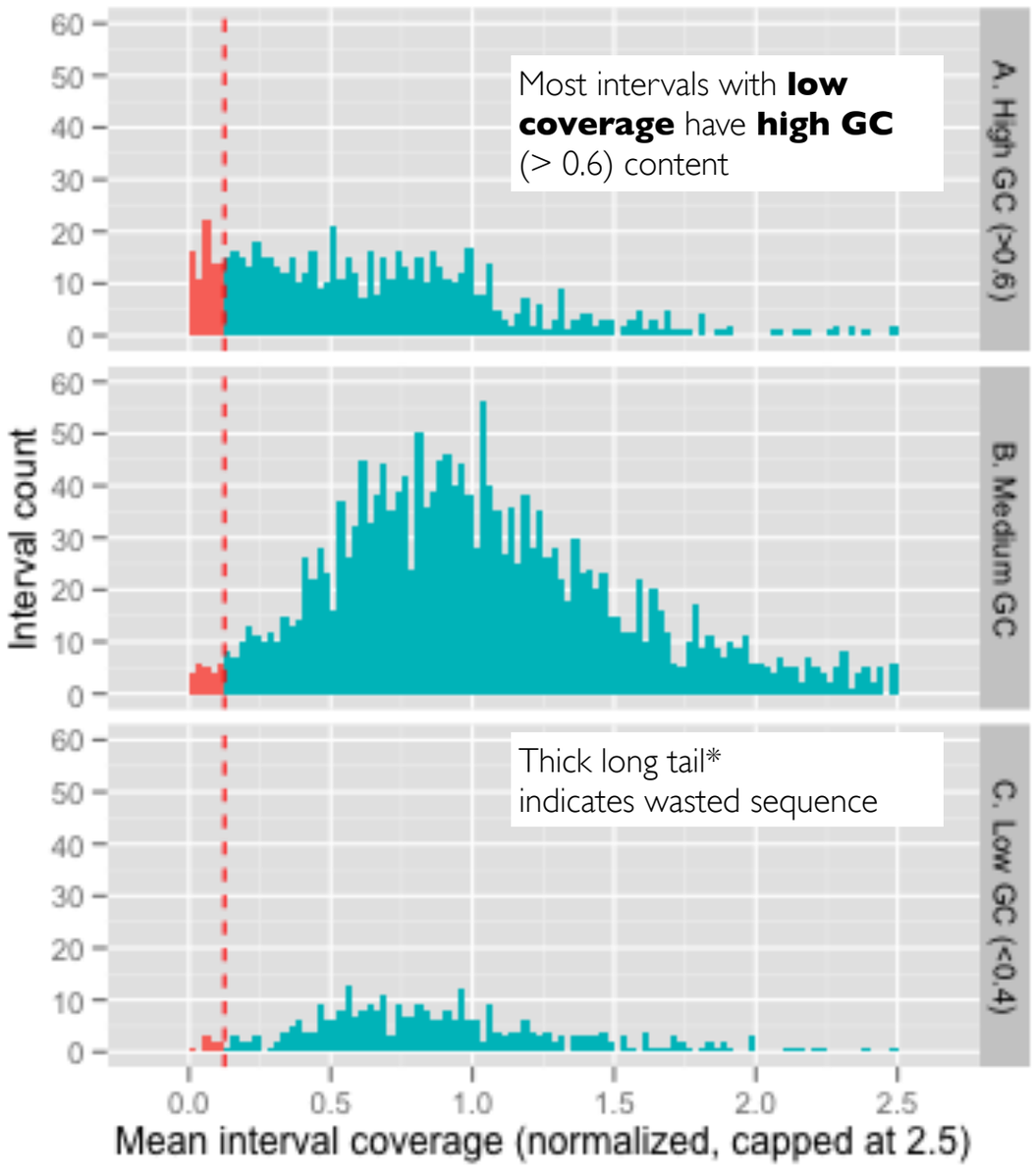


improved coverage in most low performers

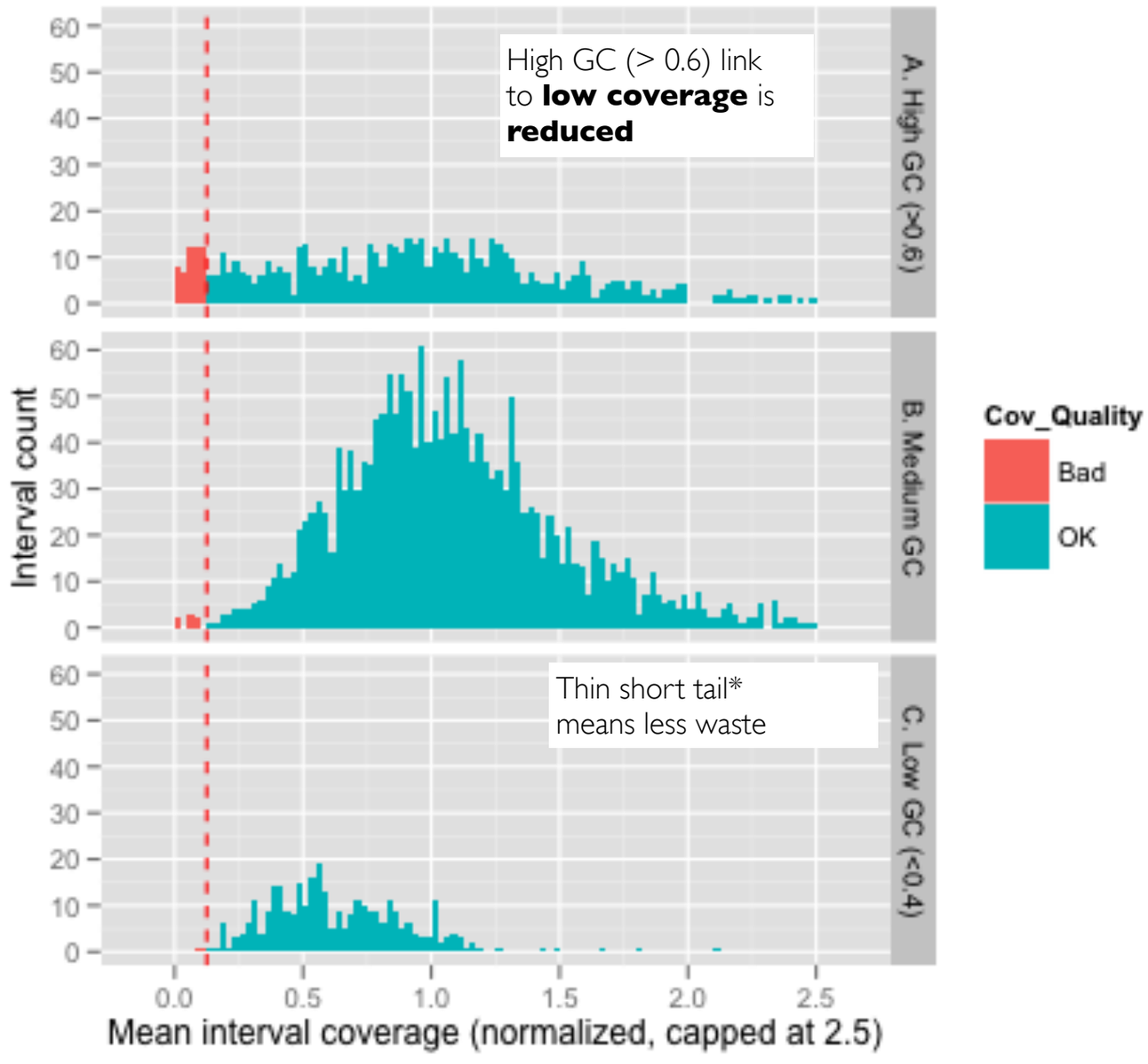
still some shared low performers

overcovered targets are now closer to the mean

Illumina's Rapid Capture Exome performs better in difficult GC content regions



current production protocol



Nextera Rapid Capture Exome

Illumina's Nextera Rapid Capture Exome also simplifies the workflow making the process faster

New Content

- Content was co-developed by Broad and Illumina
- Target regions totaling 37.7mb
- Includes:
  - All content from Broad's existing production exome
  - All coding content from the following databases as of March 2012 via the UCSC genome database:
    - CCDS (Consensus CoDing Sequence)
    - Known Gene
    - RefSeq
  - All coding content from Gencode V11

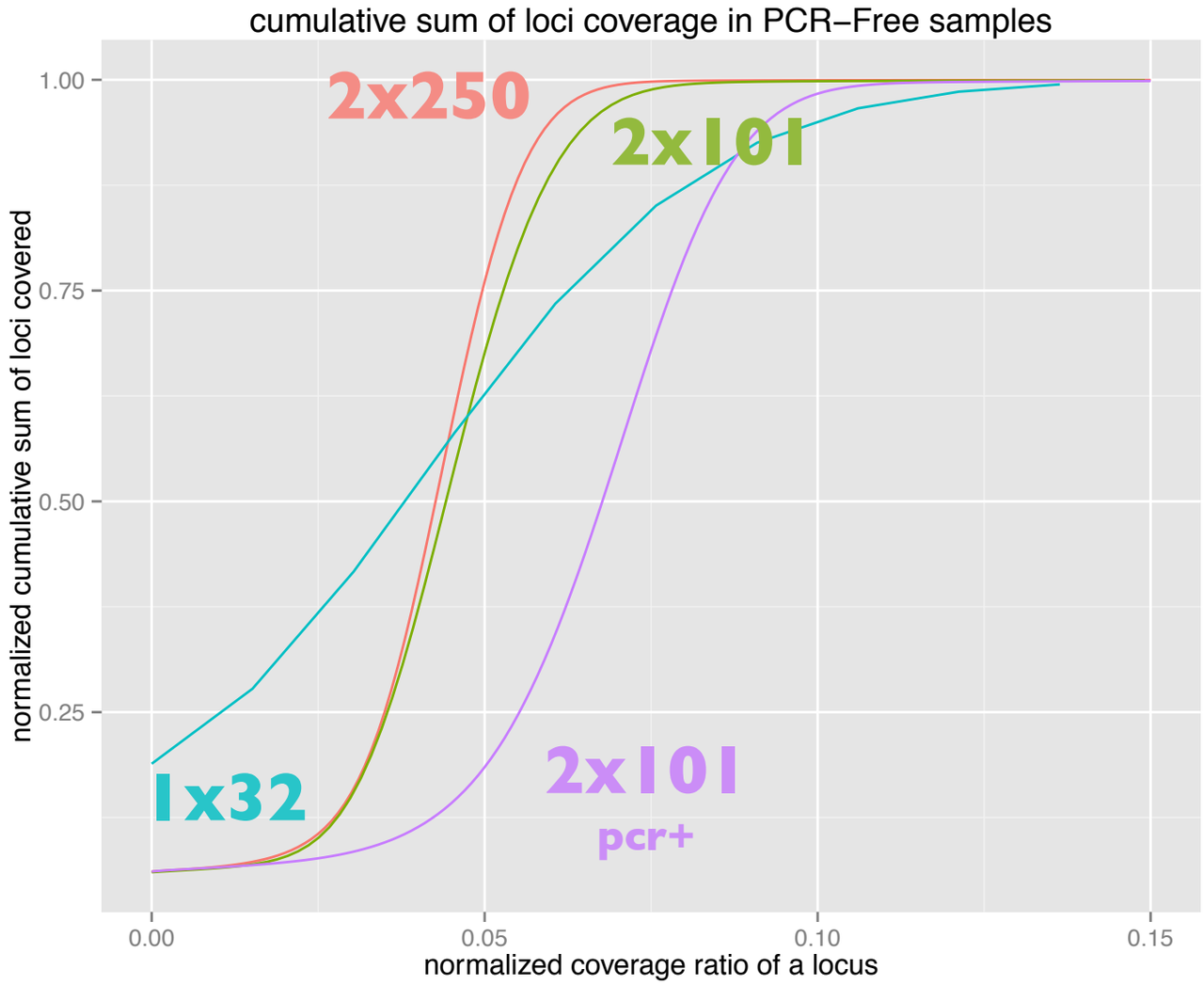
Workflow Benefits

- Low input (50ng)
- Simplified workflow
- Faster (1.5 days)
  (from genomic DNA to sequencer)
- Highly scalable
- Cost competitive

Reducing bias with longer reads

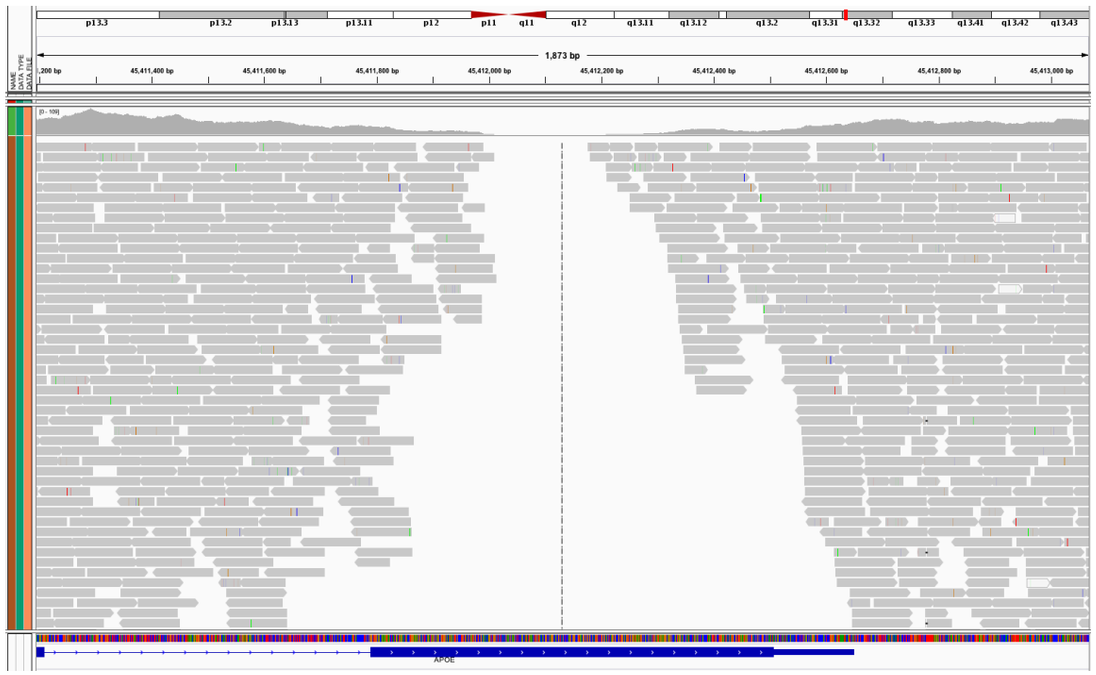
Longer reads allow us to reach further into the genome

	uncovered genome	low covered genome
2x250	5.7%	8%
2x101	6.1%	14%
1x32	18%	41%



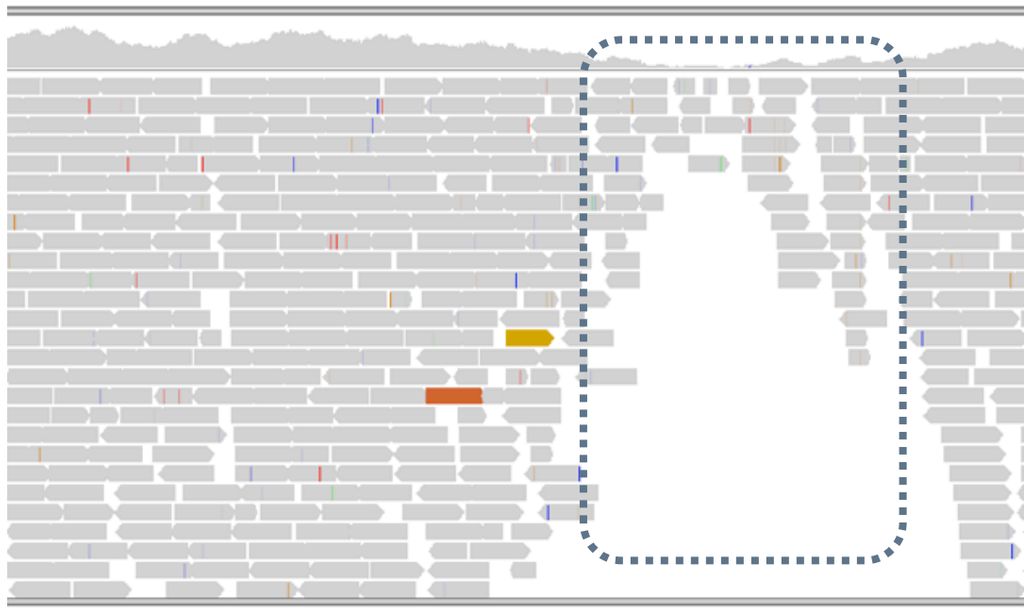
Variant calling depends on being able to see the event with confidence

APOE



completely uncovered regions are never going to get called

APOE



low covered regions will never give enough confidence to call

how much coverage do we need to call a variant?

	Sensitivity	Simple SNP	Simple INDEL	Medium INDEL	Long INDEL*
Heterozygous	50%	4x	13x	14x	13x
	90%	6x	26x	30x	26x
	99%	8x	39x	39x	32x
Homozygous	50%	2x	6x	6x	8x
	90%	2x	11x	14x	25x
	99%	2x	22x	26x	25x

76bp reads

longer reads slightly improves SNP calling

Read Length	TYPE	TRUE POSITIVES	FALSE NEGATIVES	FALSE POSITIVE BURDEN
32	SNP	269	962	0
101	SNP	1224	6	44
250	SNP	1228	2	19

this is because SNP calling is already excellent

much better indel calling with longer reads

Read Length	TYPE	TRUE POSITIVES	FALSE NEGATIVES	FALSE POSITIVE BURDEN
32	INDEL	28	251	0
101	INDEL	217	46	3
250	INDEL	232	32	1

this is because you need reads that span across the event and its surrounding repetitive context

Hot off the press: experimental 2x400 look even better

- we worked with Illumina to create a unique 2x400 WGS dataset covered to 85x.
- totally experimental sequencing tech and analytic process to evaluate the potential of even longer reads.

	uncovered genome	low covered genome
2x400	5.7%	6.1%
2x250	5.7%	8%
2x101	6.1%	14%
1x32	18%	41%

Conclusions

- For targeted sequencing, Illumina's Nextera Rapid Capture Exome product significantly reduces biases and speeds up the lab process
- For whole genome sequencing, longer reads improve coverage over previously uncovered and low covered regions of the genome.
- Calling sensitivity is improved and the specificity is highly improved by longer reads, especially for indels (the hard ones!).
- Even longer reads promise to further improve this scenario, but there is still work to do.