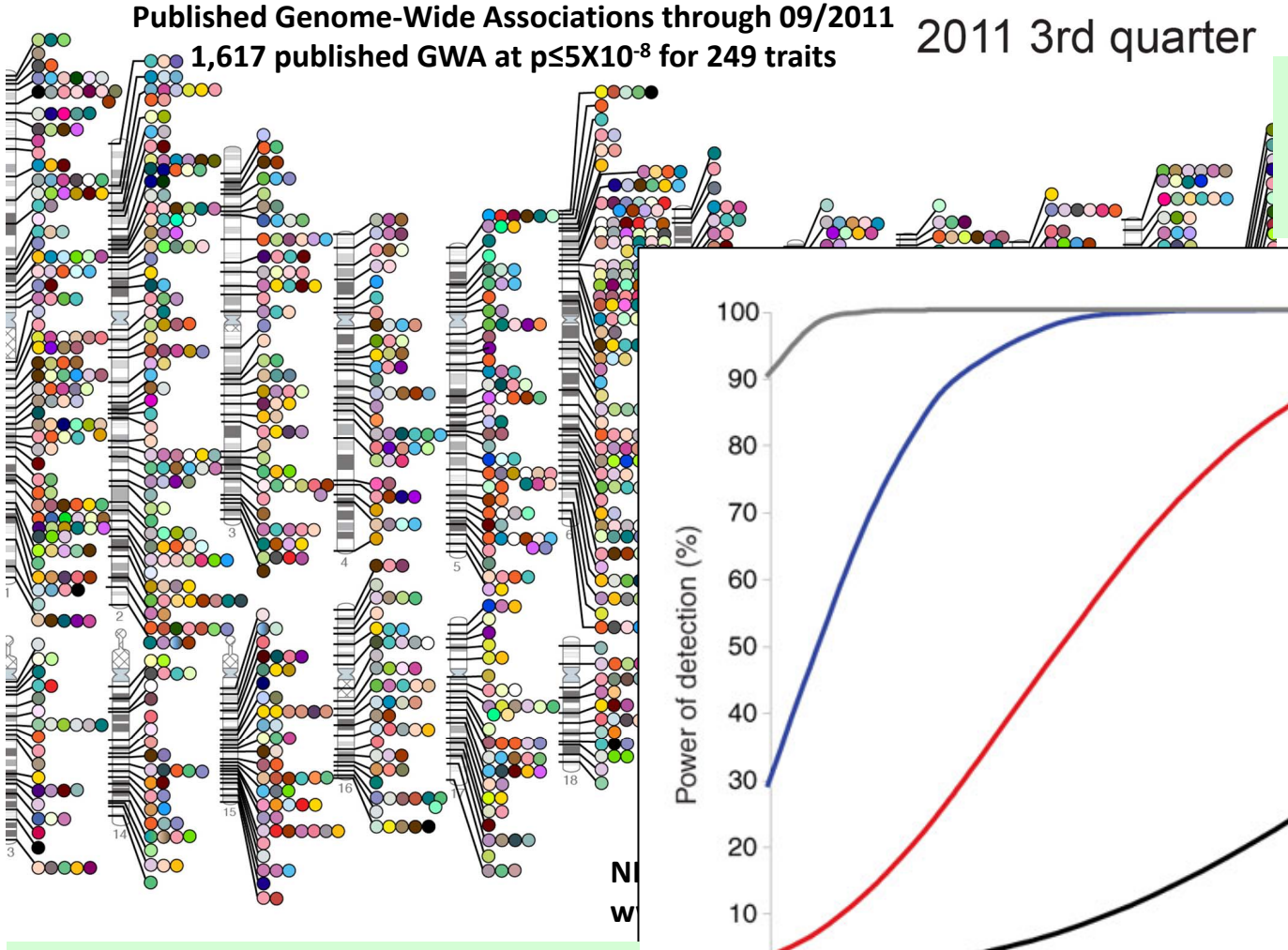
BROAD INSTITUTE

Large-Scale Genomic Variant Discovery and Validation using Pooled Sequencing Data

Guillermo del Angel, Mauricio Carneiro, Eric Banks, Ryan Poplin, Christopher Hartl, Mark DePristo
Genome Sequencing and Analysis
Medical and Population Genetics
Broad Institute of MIT and Harvard
delangel@broadinstitute.org

BROAD INSTITUTE

Challenges of a post-GWAS era

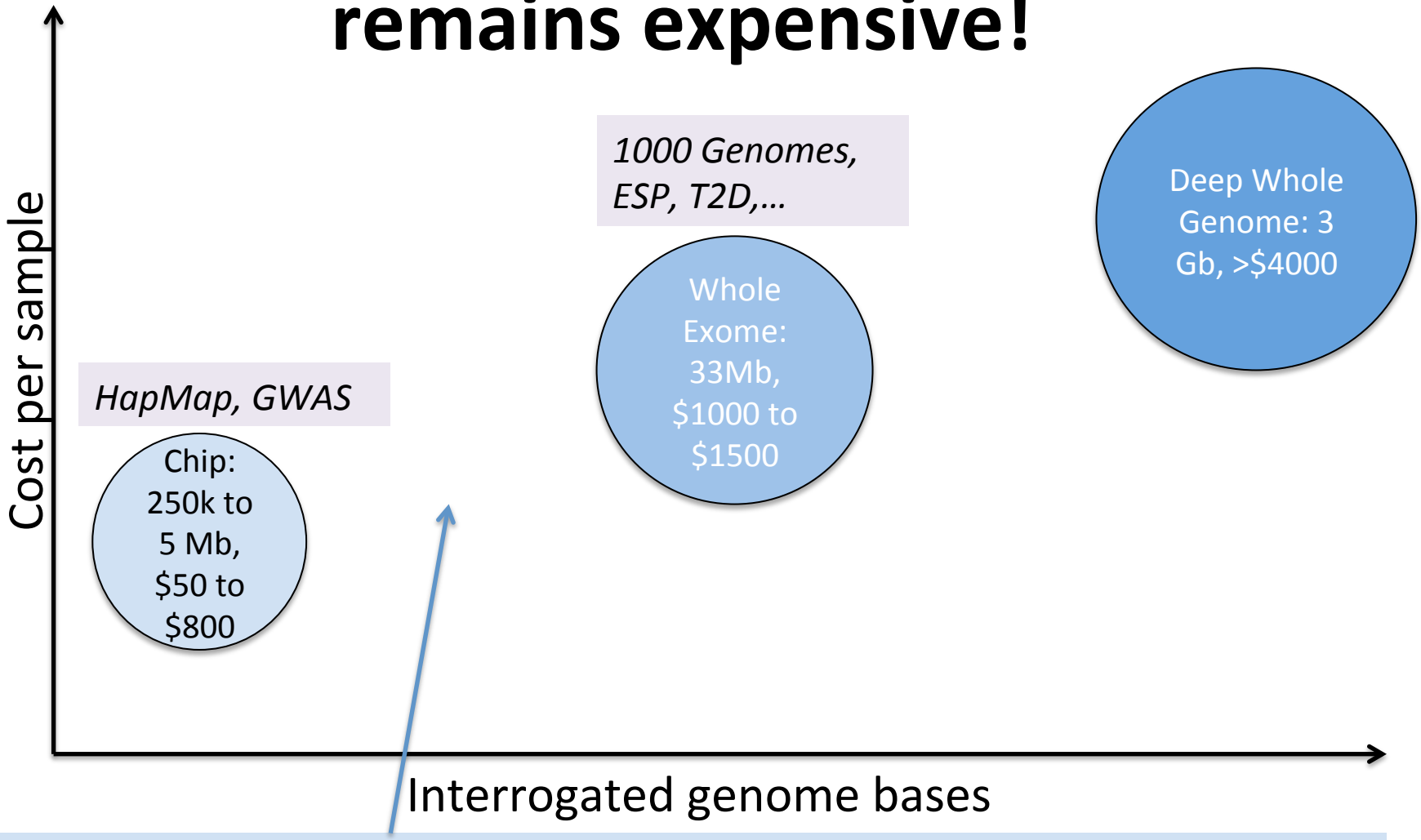


Our genetic association catalog gets bigger by the day…

… but power to probe rarer variants and smaller effects requires larger and larger sample sizes…

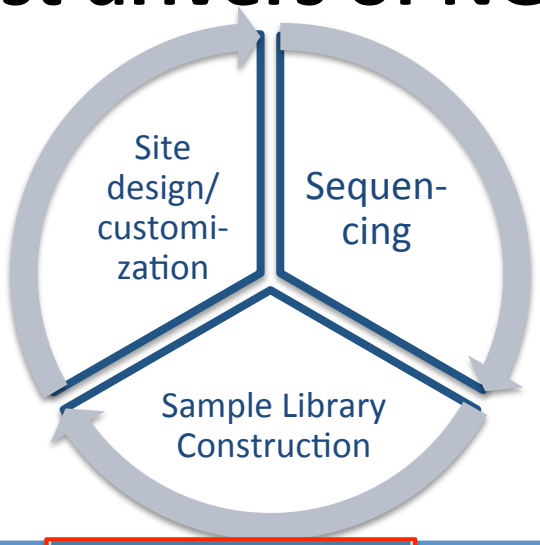
From S. Visscher et al, *Sizing up human height variation. Nat. Gen. 40, 489 (2008)*

Discovering variants in large cohorts remains expensive!



What do we do if we want to probe rarer variants in just some target regions? Variant discovery via sequencing is required for rare variant burden tests – arrays are not enough!

The three key cost drivers of NGS experiments



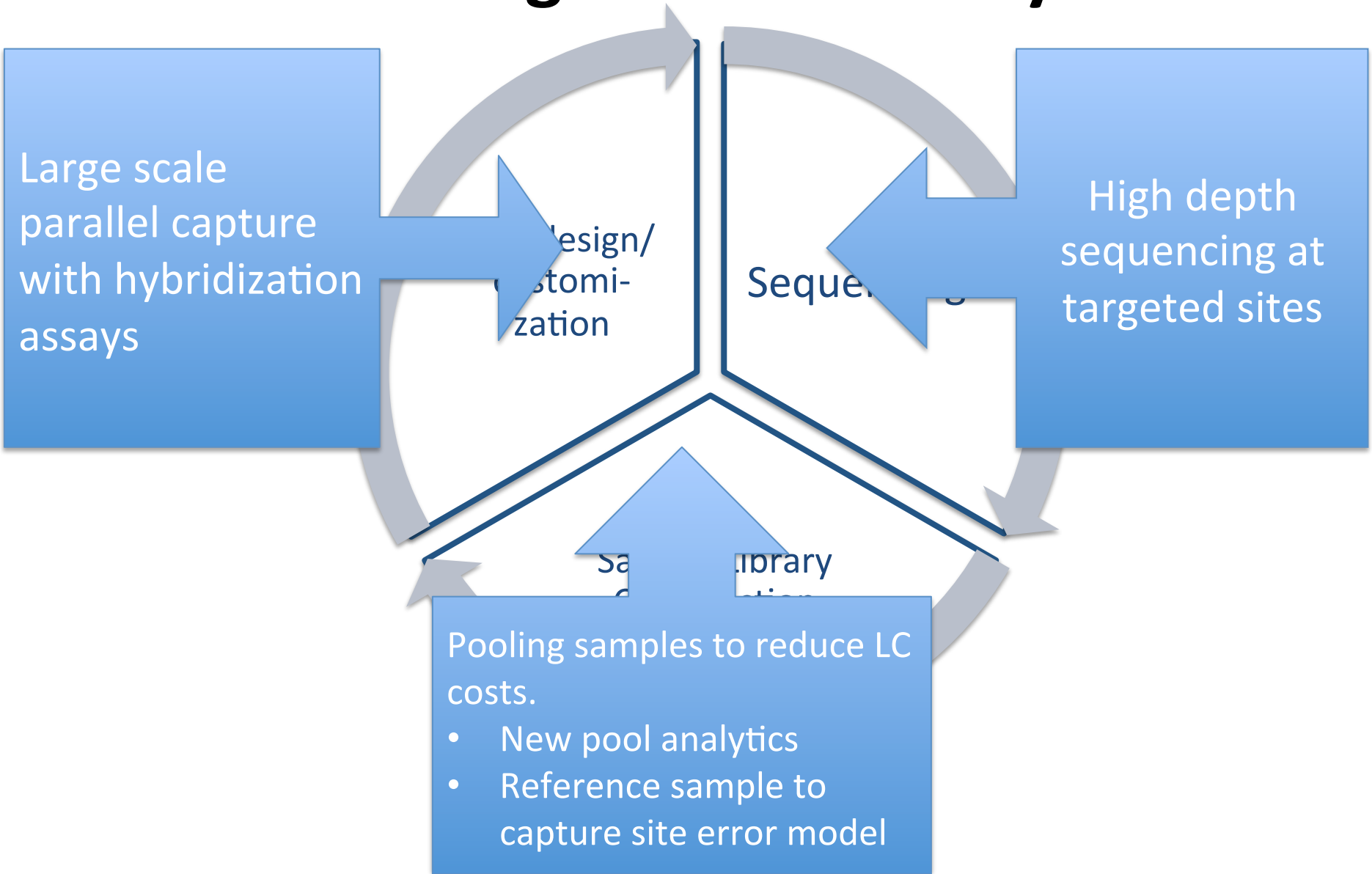
Technology	Array genotyping	Targeted capture and sequencing	Whole Exome	Whole Genome
Site Design	LOW	HIGH	LOW	NIL
Sample LC	LOW	HIGH	MEDIUM	LOW
Sequencing	NIL	LOW/MEDIUM	MEDIUM	HIGH

Sequencing costs are decreasing but other costs are lowering at a slower rate:

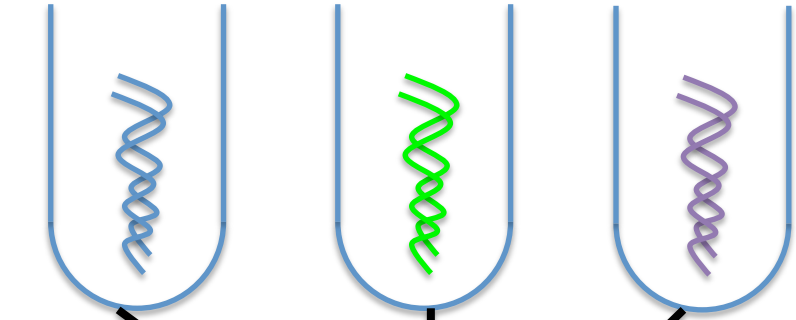
- Whole exome now costs ~$1,000, but a library creation can cost up to $400 per sample

We need new methods for assessing and discovering variants at a large scale.

Our approach deals with these challenges in three ways

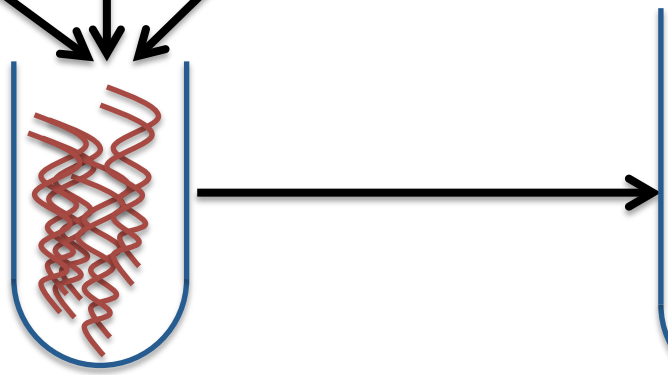


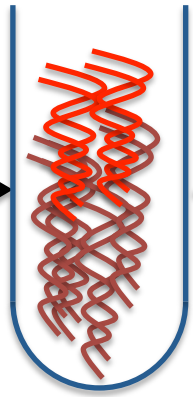
We address the challenges of sample pooling by including a bar-coded reference sample to be sequenced jointly



Typical Pooling drawbacks:
- Analytics become harder
- Sensitive to pool imbalances
- Hard to estimate error process

Pooling and barcoding each pool





Barcoded Reference sample added at 10% dilution

Capture and sequencing

Presence of reference sample allows us to estimate site error properties accurately

Traditional calling approach

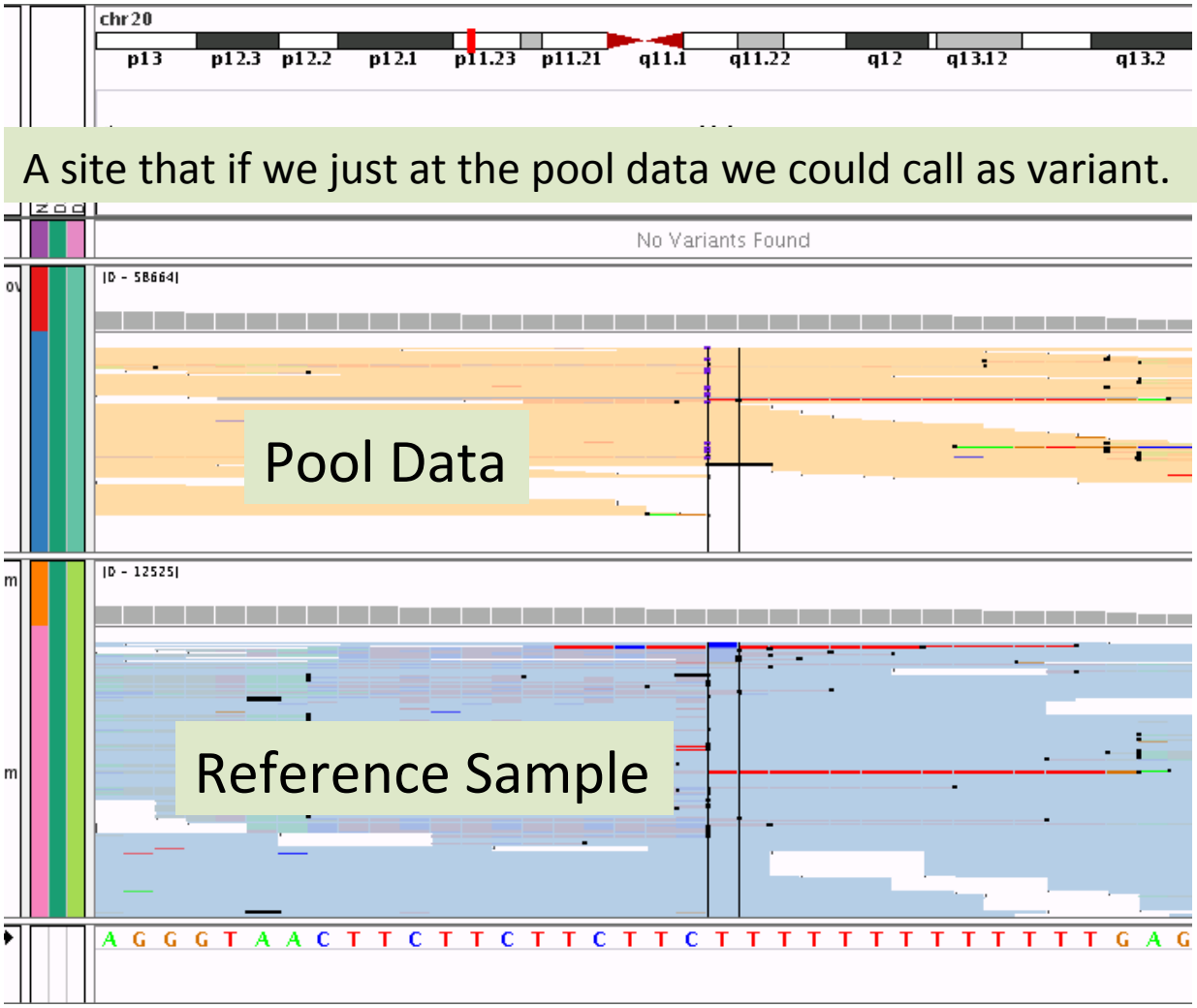
- Only base qualities (raw or recalibrated) used for statistical determination of genotype likelihoods.
- Site-dependent error properties not explicitly modeled.

Reference sample-guided approach

- For every site of interest, get “truth” genotypes from reference sample (NA12878 in our case).
- Site-dependent error properties are captured by scoring actual reference sample sequence data with truth genotypes.

Reference sample-guided variant calling is not exclusive to pool calling!

Errorful sites are thus removed a priori from callset



Availability of pre-existing high-quality "truth" genotypes for NA12878 allows us to build a statistical model for each site of interest. Systematic sequencing errors or artifacts are then eliminated because statistical evidence for variant is adjusted accordingly.

We can target and enrich large numbers of genomic regions simultaneously and sequence pools to validate large number of variants

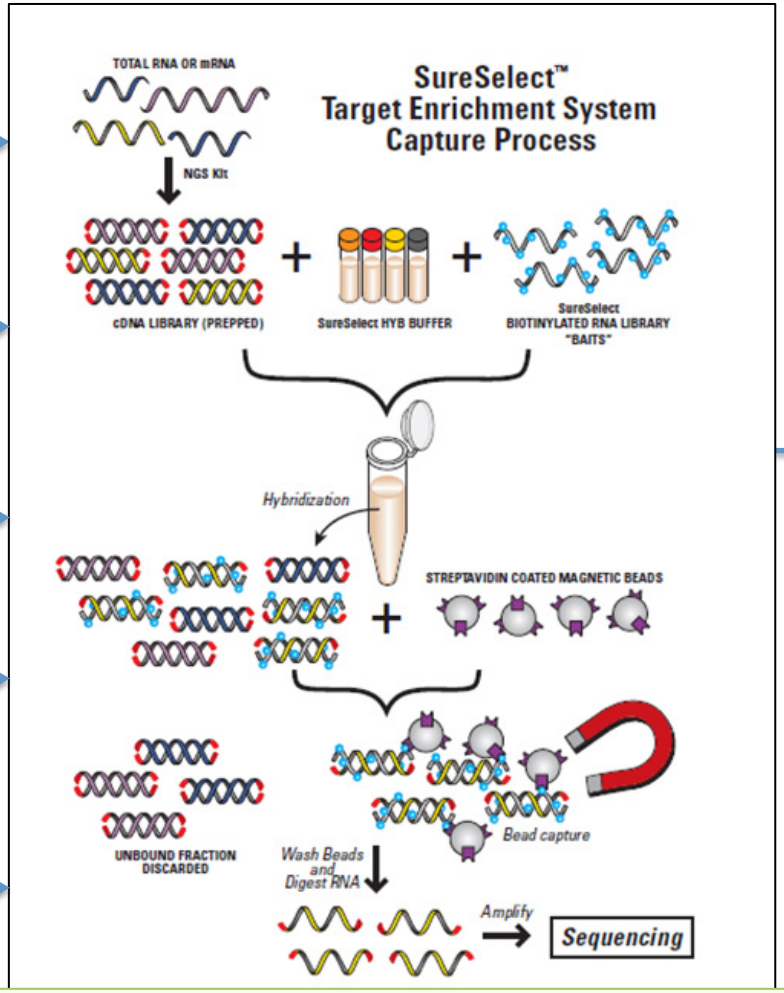
Pooled/bar-coded samples

Pooled/bar-coded samples

Pooled/bar-coded samples

Bar-coded reference sample

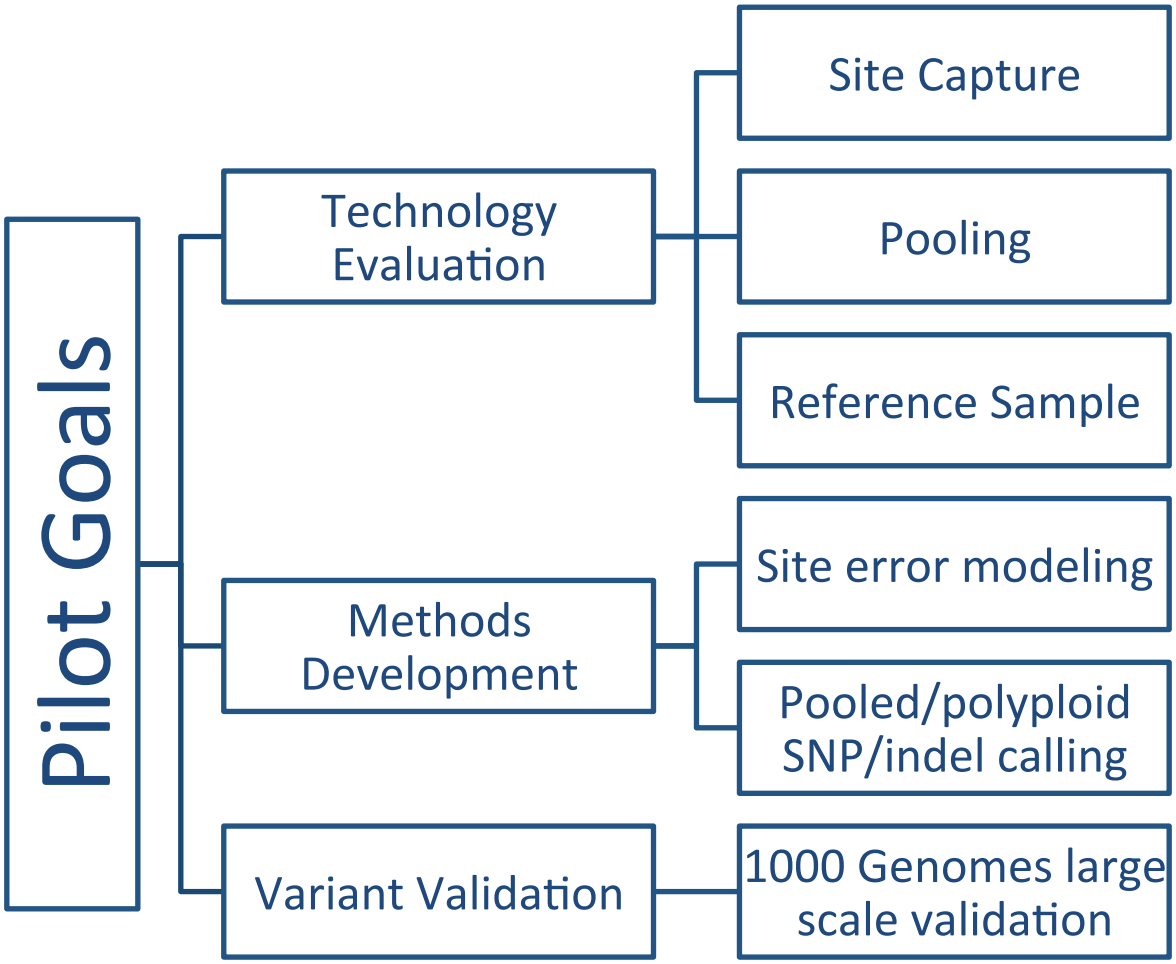
Pooled/bar-coded samples



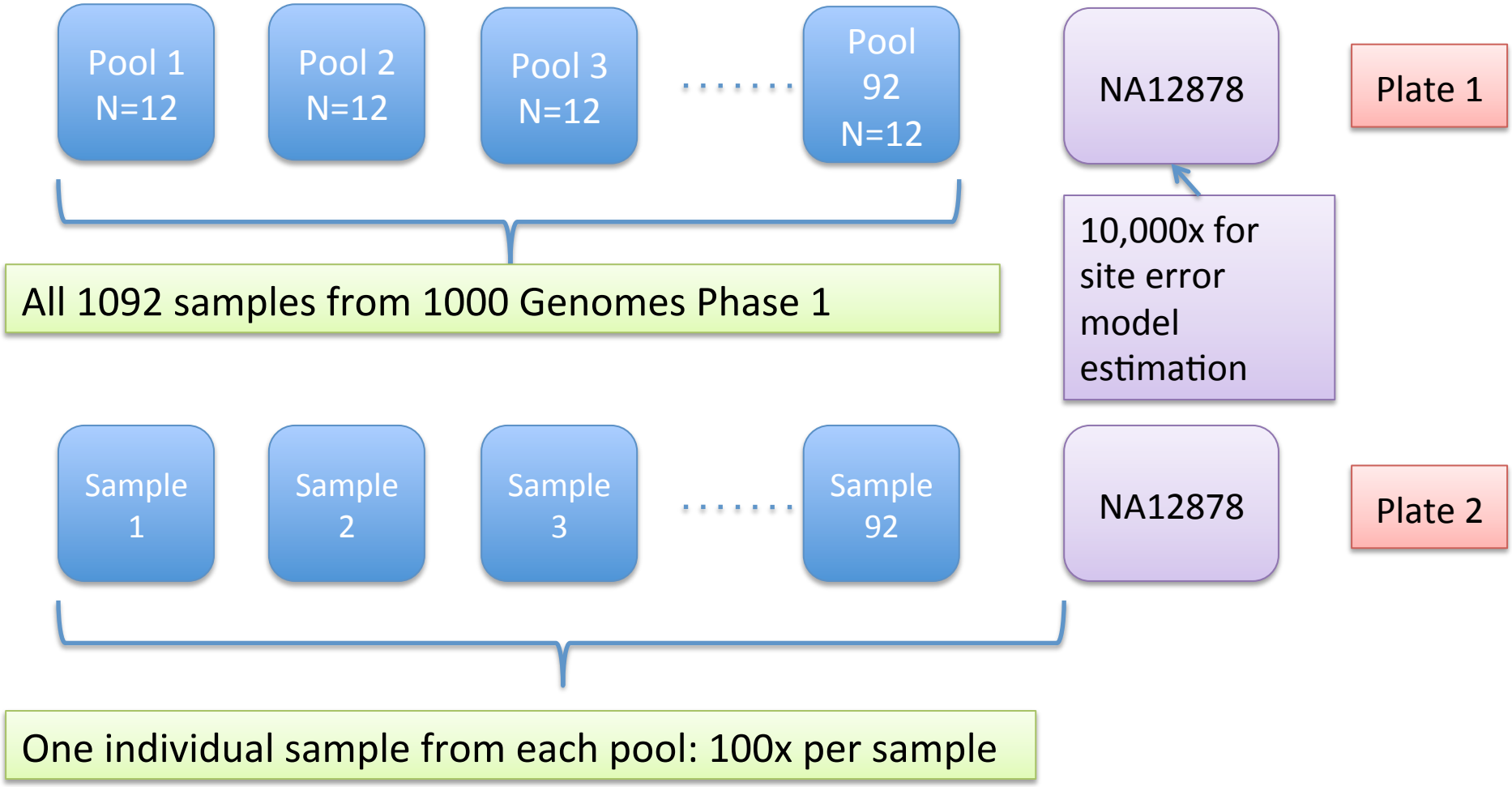
Deep Sequencing

Baits designed around sites of interest.

We performed a pilot experiment to prototype this technology

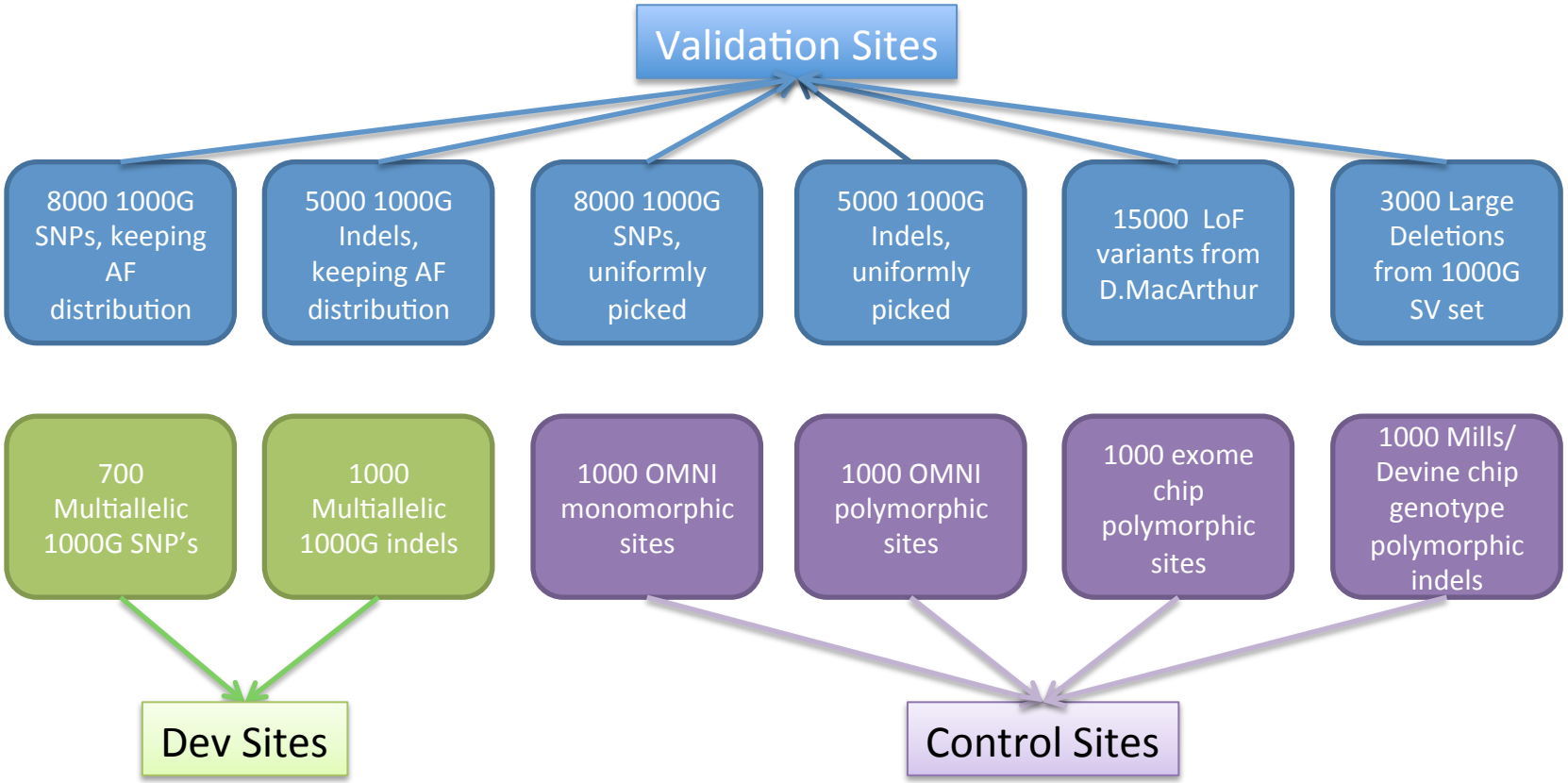


Experimental design: 50,000 sites @ 1200x each pool



Barcoded NA12878 added at target 10 % dilution to each pool. Yields aggregate target NA12878 depth of about 10,000x per validation site

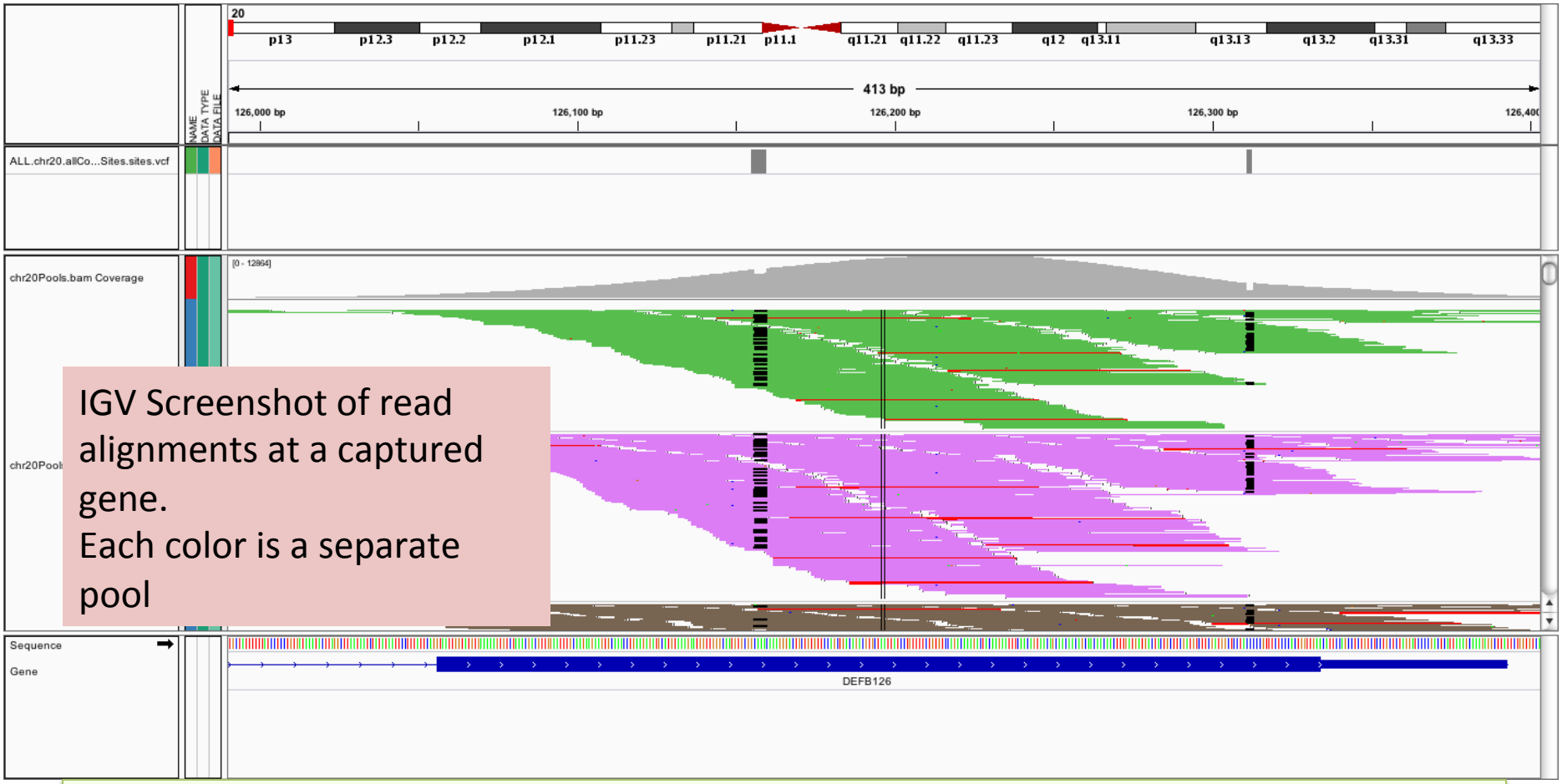
Validation Site Design



- SNPs and indels in large 1000G Phase 1 sets were picked if they were polymorphic in 8 validation samples.
- LoF Variants are SNPs and indels.
- Phase 1 indels chosen from the pre-SVM filtered set.
- Large deletion set consists of 2700 probes for flank and 400 probes for alt sequence.

Control sites chosen to assess accuracy of capture and calling mechanisms

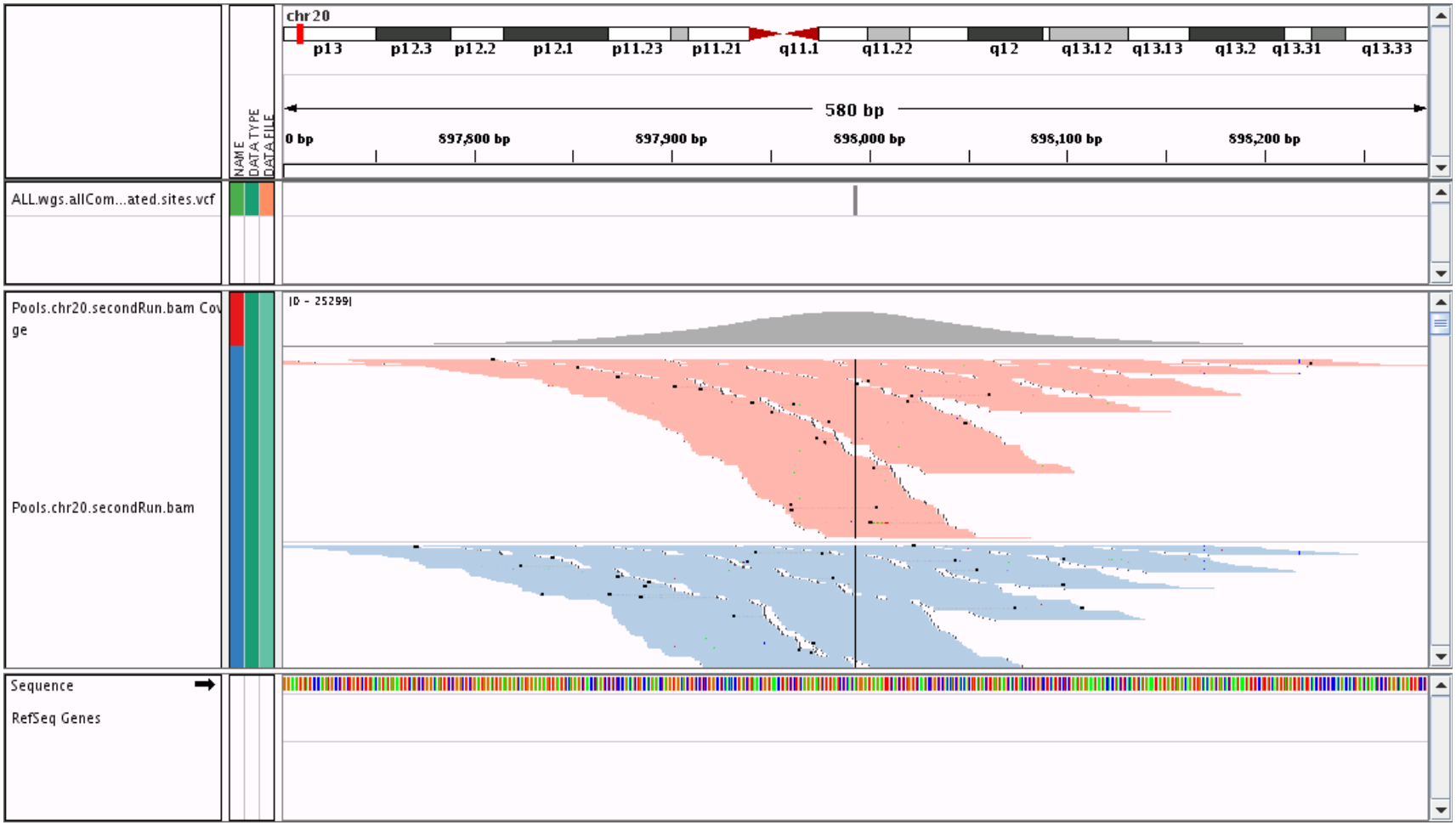
Resulting reads show successful capture and sequencing around targeted variant sites



IGV Screenshot of read alignments at a captured gene.
Each color is a separate pool

Two LOF indels clearly present in many of the pools.
Successful sequencing of 90/92 pools in ~48,500 baits.

We also find some sites which are false positives in 1000 Genomes



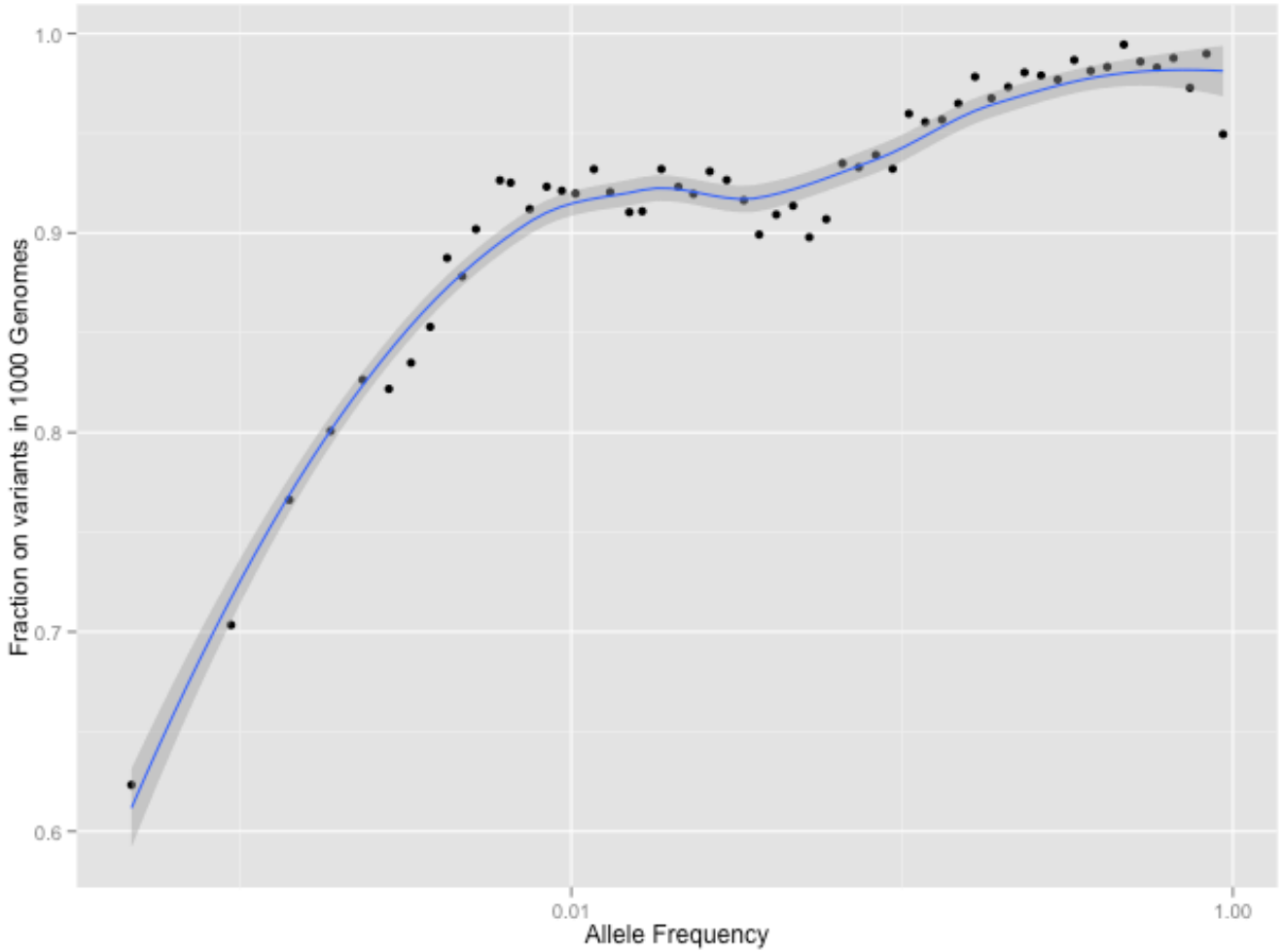
A SNP from 1000G with no apparent support in reads

Calls at control sites show that we can clearly discriminate true and false variation

Set	Pool Caller called Monomorphic (AC=0)	Pool Caller called Polymorphic (AC>0)	No-call/Filtered (not enough coverage)
OMNI Mono (SNPs)	711	168	101
OMNI Poly (SNPs)	6	956	38
Exome Chip (SNPs)	3	956	41
Mills Indel Chip	14	940	46

3 Exome Chip SNP sites called monomorphic shows that caller is doing what it's expected to do: no evidence of polymorphism in 1000G samples.

Well over 90 % of all discovered SNPs with AF > 1% are already in 1000 Genomes Phase 1



85,159 SNPs called in all designed baits and filtered by standard VQSR and depth. Missing low frequency variants are a combination of false positives in pool caller and lower sensitivity of 1000 Genomes low-pass sequencing

1000 Genomes SNP and Indel site validation consistent with published rates

Data Set	# Called sites [1]	FDR (%)
AF SNPs	6166	1.9 %
AF Indels, post-SVM filtering	1326	18.0 %
AF Indels, pre-SVM filtering [3]	3591	39.2 %
LOF SNPs	5207	5.7 %

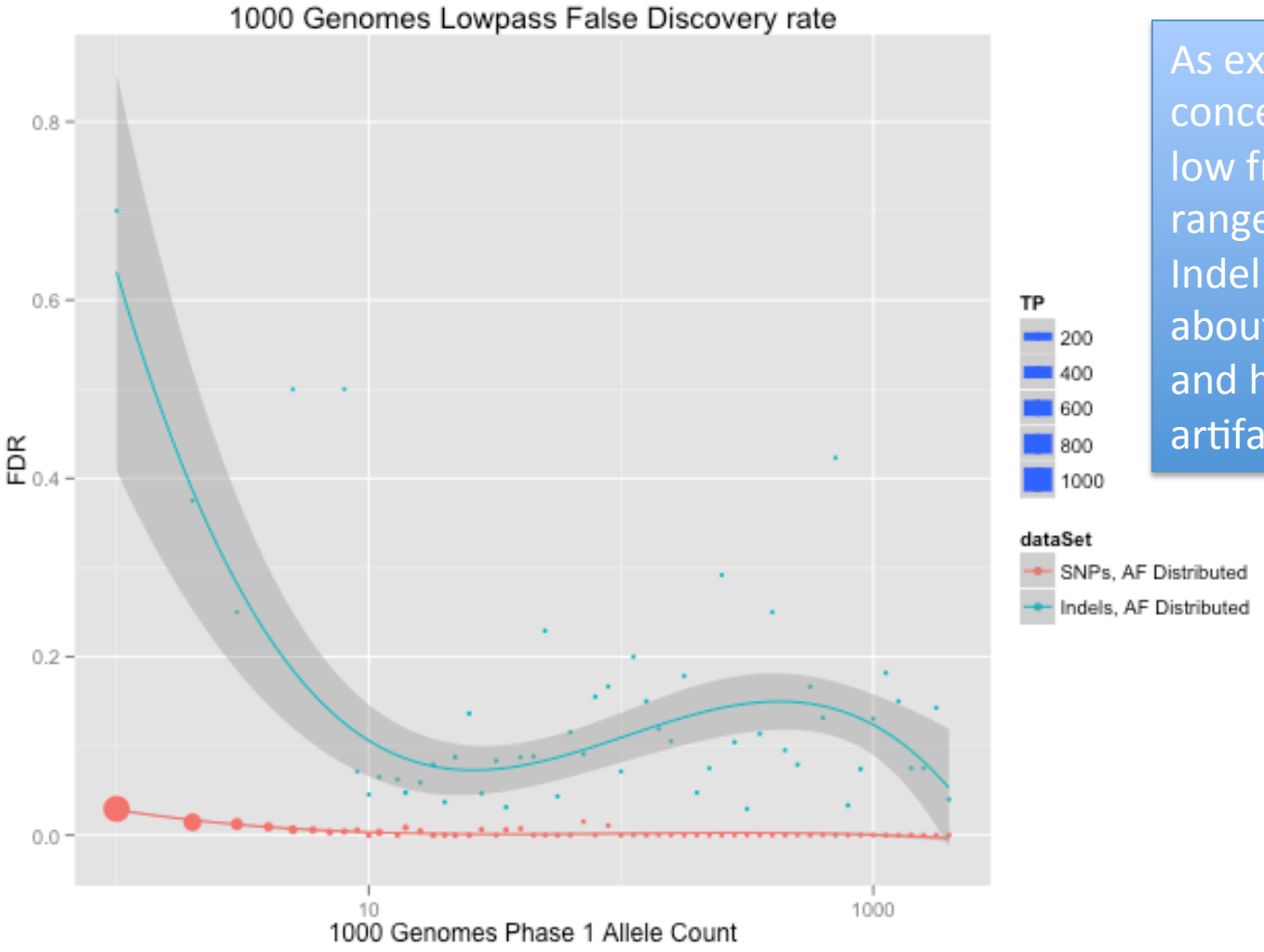
NOTES:

1. Only validation sites that had total depth > 5000 and Reference Sample Depth > 500 were kept.
2. Bait design and site selection were done after preliminary 1000G integration was done, but before final SVM filtering removed many indels.

We can call variants in about 70-80% of the sites, but filtering could be relaxed to recover more sites.

Lowpass FDR published in [Nature, 2012] paper: 1.8 % (SNPs), 35.5 % (Indels, pre-filtering)

Large number of sites allows us to compare errors across AF spectrum



As expected, FDR concentrated on low frequency range.
Indel FDR is still about 10x SNP FDR and high-frequency artifacts remain

Conclusions

- Novel approach of combining targeted capture, high depth sequencing with pooling and addition of reference sample to capture site error modes allows us to perform very accurate discovery and validation experiments.
- This approach is being prototyped for several projects and is under active development and improvement.
- Future work will involve applying this methodology to large-scale clinical sequencing experiments.
- Methodology is also being continually updated in the GATK framework.
- We intend to perform another round of large-scale validation for 1000 Genomes using this methodology for Phase 3.

Acknowledgements

Broad Institute

- **David Altschuler**
- Eric Banks
- Mauricio Carneiro
- **Mark DePristo**
- Yossi Farjoun
- Sheila Fisher
- **Stacey Gabriel**
- Namrata Gupta
- Bob Handsaker
- Heng Li
- Daniel MacArthur
- April Monchik
- Ryan Poplin
- David Roazen
- Khalid Shakir
- Geraldine van der Auwera

1000 Genomes Project

- Goncalo Abecasis
- Danny Challis
- Laura Clarke
- Scott Devine
- Richard Durbin
- Erik Garrison
- Hyun Min Kang
- Gil McVean
- **… and the rest of the Analysis Group!**