Large-Scale Variant Validation using Pooled Sequencing

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

Summary

- We're presenting the results of our Large Scale validation experiment on all 1092 Phase 1 samples.
- We chose ~50,000 SNP+Indel+Large Deletion sites, got validation data on about 40,000 passing sites.
- SNP and Indel Validation rates mostly in line with published results in *Nature* paper.
- We have a wealth of new information that we can leverage to improve our calling methods.

We’ve learned a lot on how to call and validate variants, but we have ways to go

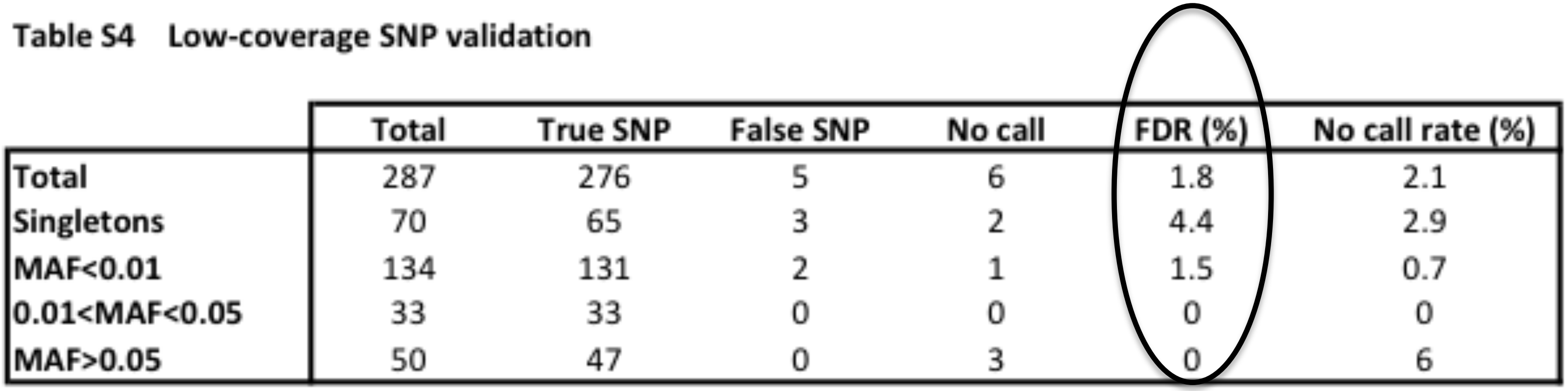


Table S4 Low-coverage SNP validation

	Total	True SNP	False SNP	No call	FDR (%)	No call rate (%)
Total	287	276	5	6	1.8	2.1
Singletons	70	65	3	2	4.4	2.9
MAF<0.01	134	131	2	1	1.5	0.7
0.01<MAF<0.05	33	33	0	0	0	0
MAF>0.05	50	47	0	3	0	6

Low-coverage Indel Validation from 1000 Genomes showed about 20x higher FDR than SNPs!

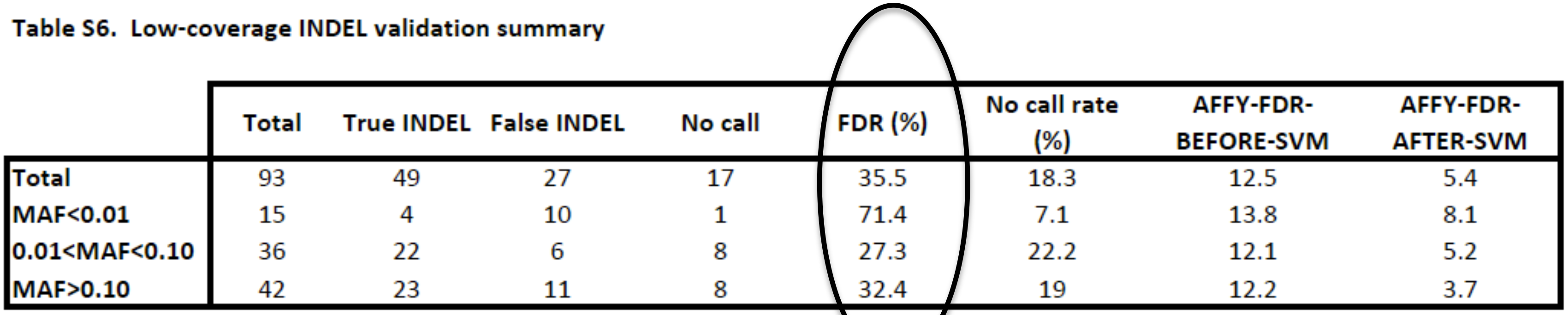


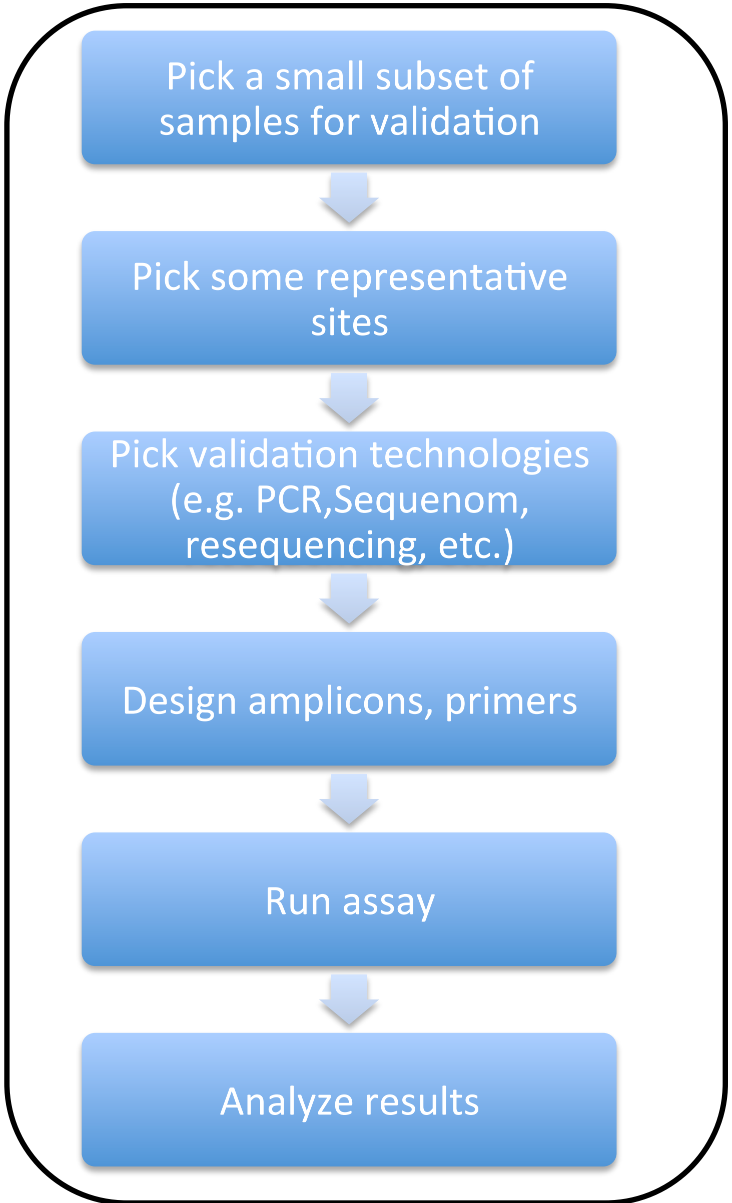
Table S6. Low-coverage INDEL validation summary

	Total	True INDEL	False INDEL	No call	FDR (%)	No call rate (%)	AFFY-FDR-BEFORE-SVM	AFFY-FDR-AFTER-SVM
Total	93	49	27	17	35.5	18.3	12.5	5.4
MAF<0.01	15	4	10	1	71.4	7.1	13.8	8.1
0.01<MAF<0.10	36	22	6	8	27.3	22.2	12.1	5.2
MAF>0.10	42	23	11	8	32.4	19	12.2	3.7

From *“An Integrated Map of genetic variation from 1092 Genomes”*, Nature, in print

Traditional validation methods don't scale when assessing accuracy of large datasets

Traditional Validation Workflow



Pick a small subset of samples for validation

Pick some representative sites

Pick validation technologies (e.g. PCR,Sequenom, resequencing, etc.)

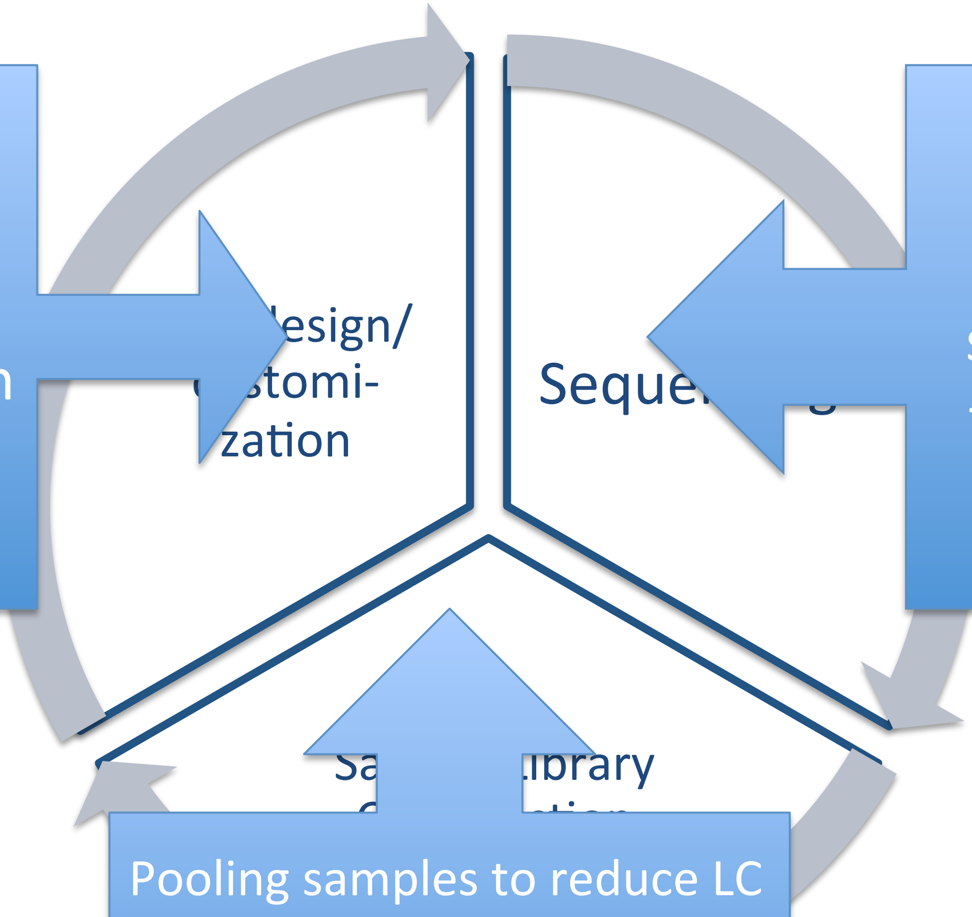
Design amplicons, primers

Run assay

Analyze results

- Validation is hard!
  - Validation discordance among multiple technologies.
  - Error modes particular to technologies.
  - Validating in a small subset of samples conflates genotyping and site discovery issues.
  - Need large number of genomic sites to assess accuracy
- Sequencing is getting cheaper quickly but library creating isn't!
  - Per-sample preparation cost may dominate validation budget

We've developed an approach that deals with some of these challenges in three ways



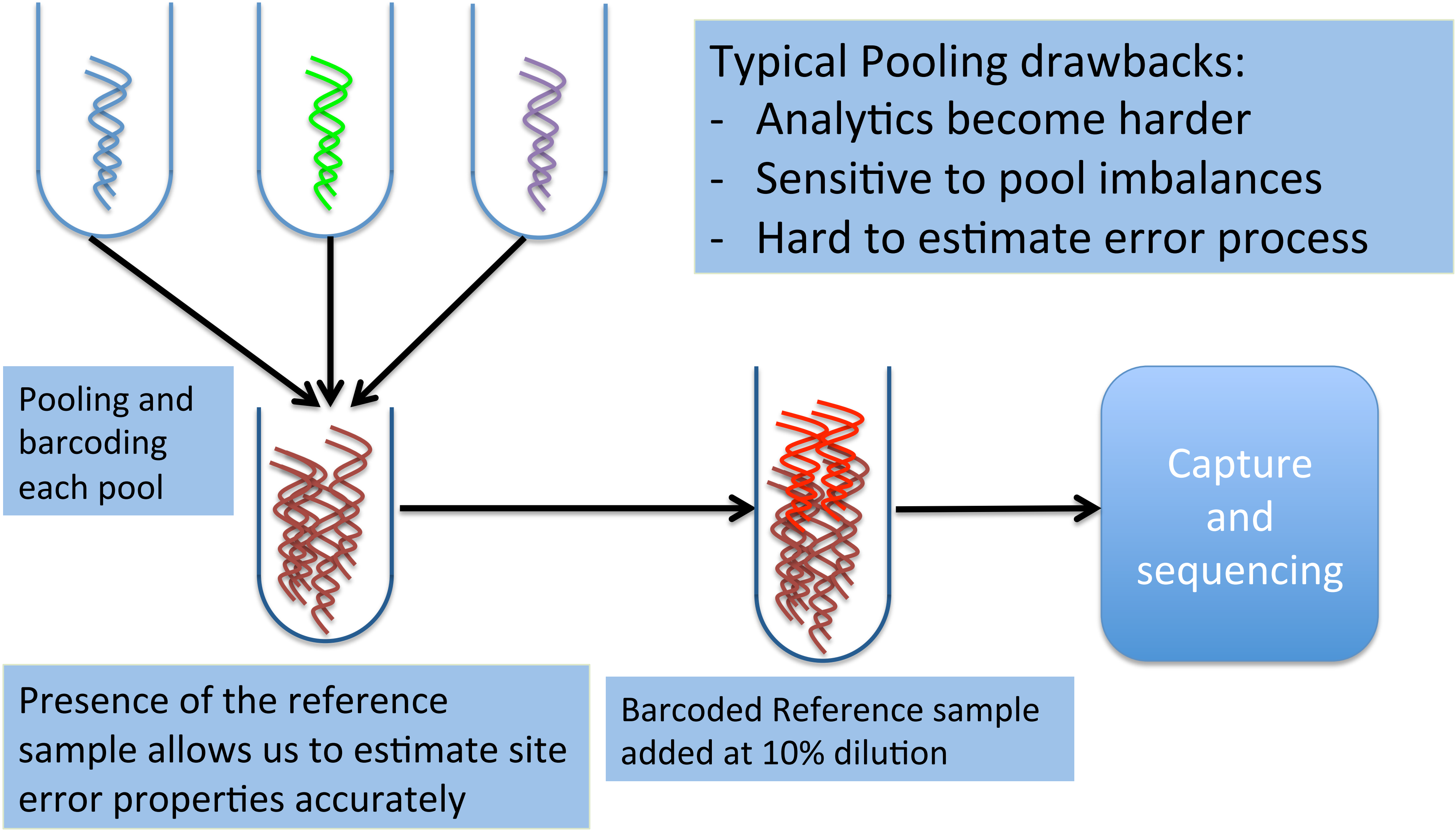
Large scale parallel capture with hybridization assays

High depth sequencing at targeted sites

Pooling samples to reduce LC costs.

- New pool analytics
- Reference sample to capture site error model

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

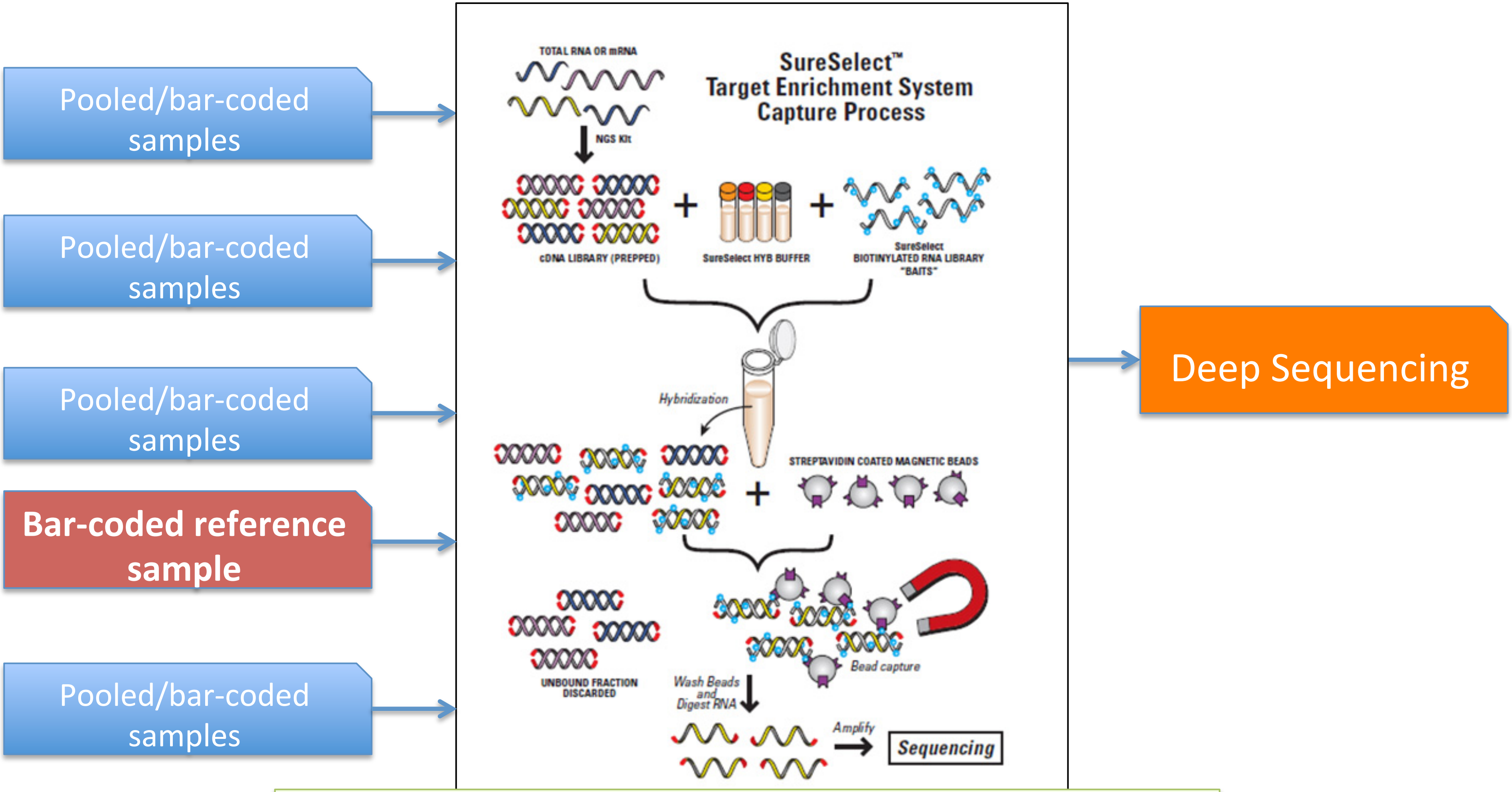
Pooling and barcoding each pool

Barcoded Reference sample added at 10% dilution

Capture and sequencing

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

We targeted and enriched large numbers of genomic regions simultaneously and sequenced pools to validate ~50,000 variants



Pooled/bar-coded samples

Pooled/bar-coded samples

Pooled/bar-coded samples

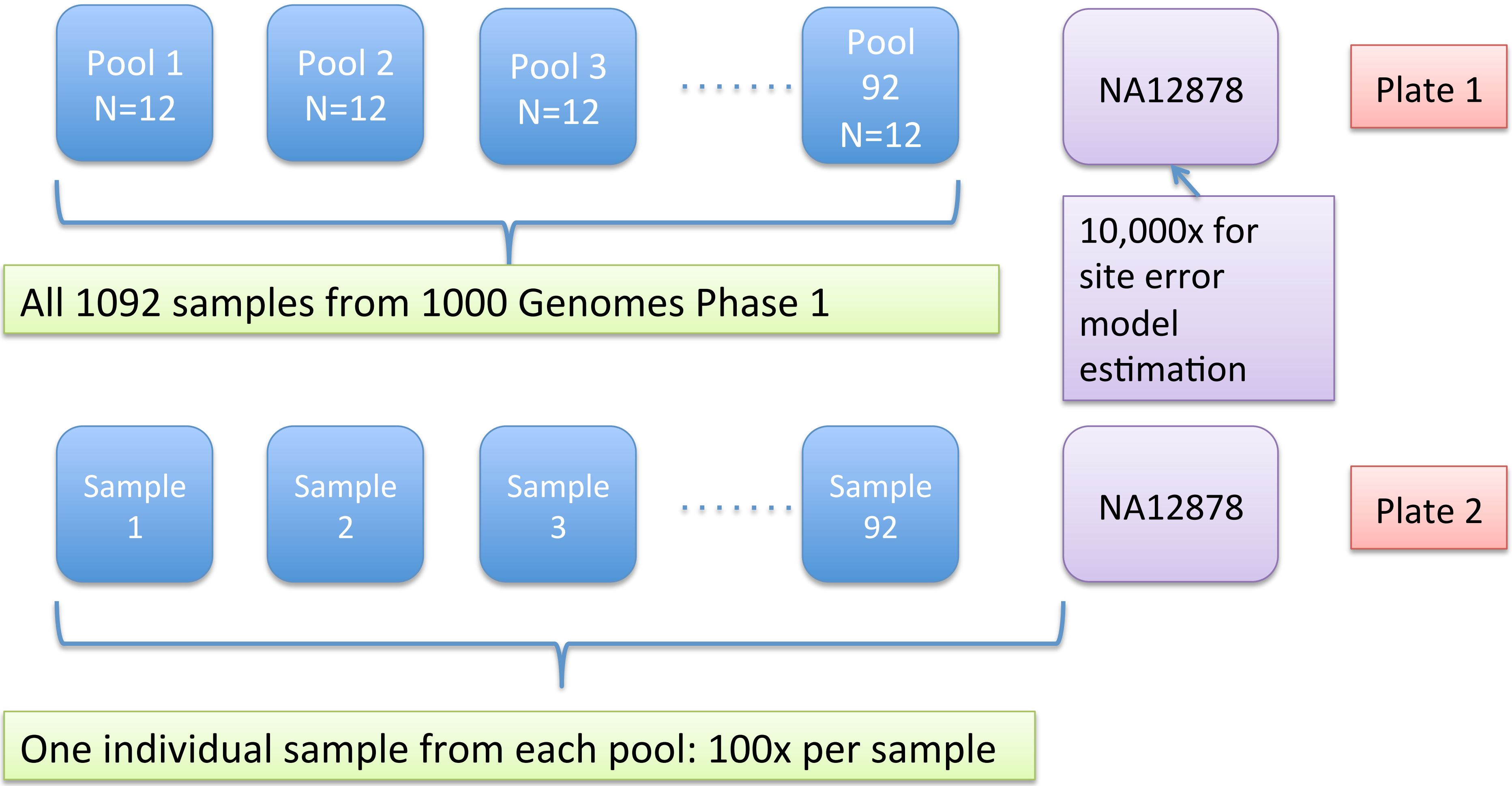
Bar-coded reference sample

Pooled/bar-coded samples

Deep Sequencing

Baits designed around sites of interest.

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



Pool 1 N=12

Pool 2 N=12

Pool 3 N=12

. . . .

Pool 92 N=12

NA12878

Plate 1

10,000x for site error model estimation

All 1092 samples from 1000 Genomes Phase 1

Sample 1

Sample 2

Sample 3

. . . .

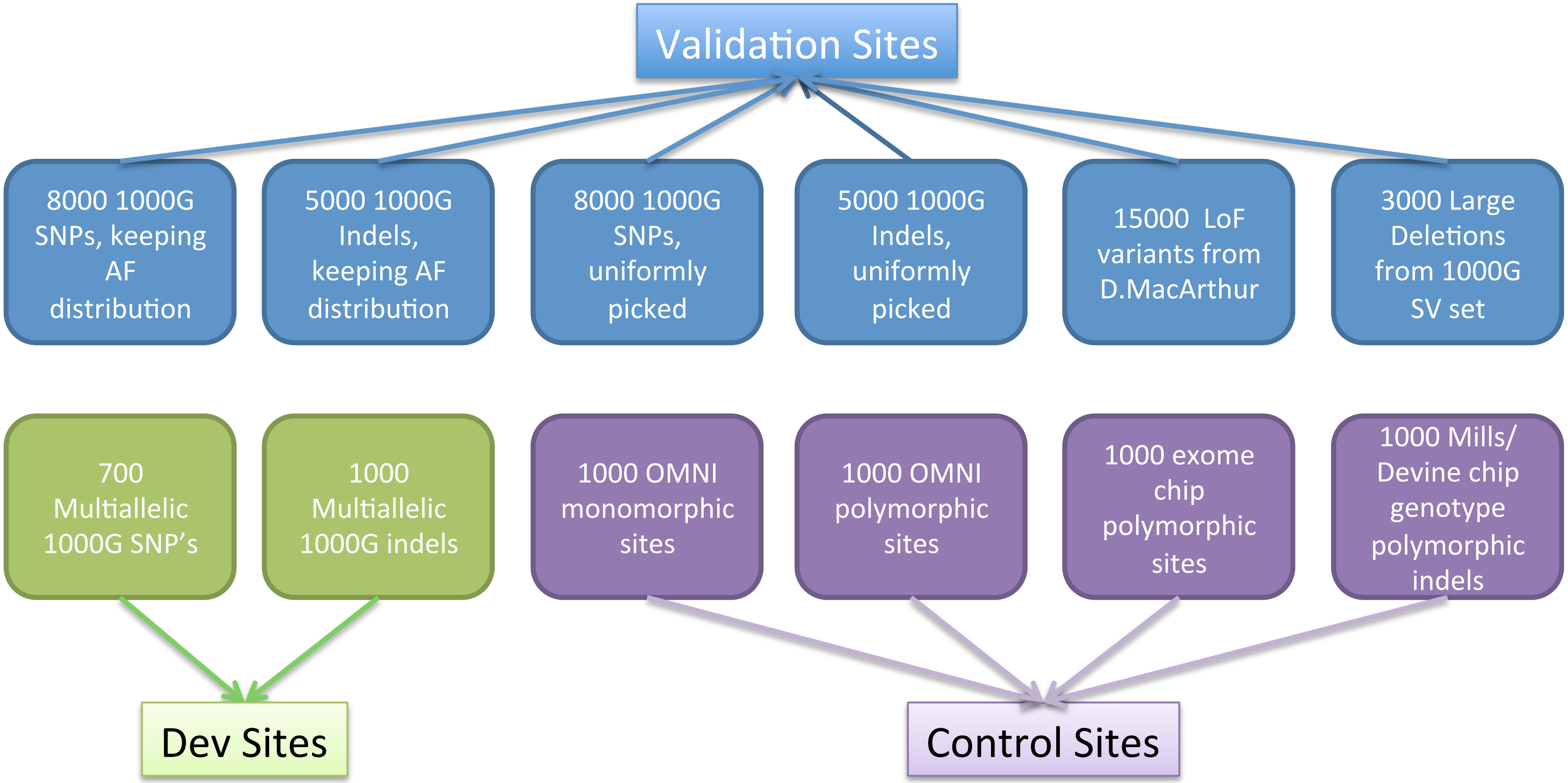
Sample 92

NA12878

Plate 2

One individual sample from each pool: 100x per sample

Validation Site Design



Validation Sites

8000 1000G SNPs, keeping AF distribution

5000 1000G Indels, keeping AF distribution

8000 1000G SNPs, uniformly picked

5000 1000G Indels, uniformly picked

15000 LoF variants from D.MacArthur

3000 Large Deletions from 1000G SV set

700 Multiallelic 1000G SNP's

1000 Multiallelic 1000G indels

1000 OMNI monomorphic sites

1000 OMNI polymorphic sites

1000 exome chip polymorphic sites

1000 Mills/ Devine chip genotype polymorphic indels

Dev Sites

Control Sites

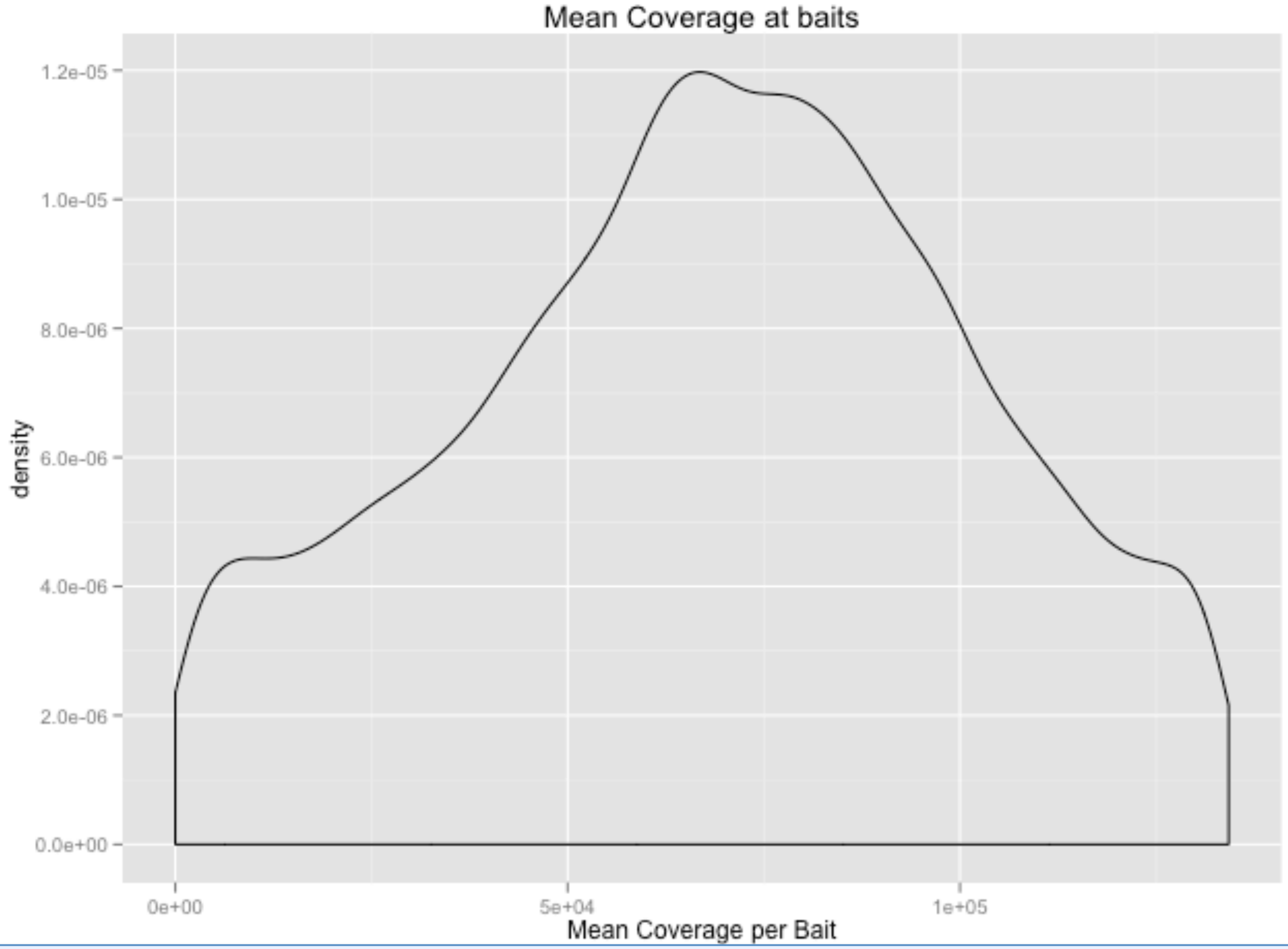
- SNPs and indels in large 1000G Phase 1 sets were picked if they were polymorphic in 8 validation samples.
- LoF Variants are SNPs and indels.
- Phase 1 indels chosen from the pre-SVM filtered set.
- Large deletion set consists of 2700 probes for flank and 400 probes for alt sequence.

Control sites chosen to assess accuracy of capture and calling mechanisms

Basic metrics show successful sequencing of pools and individuals

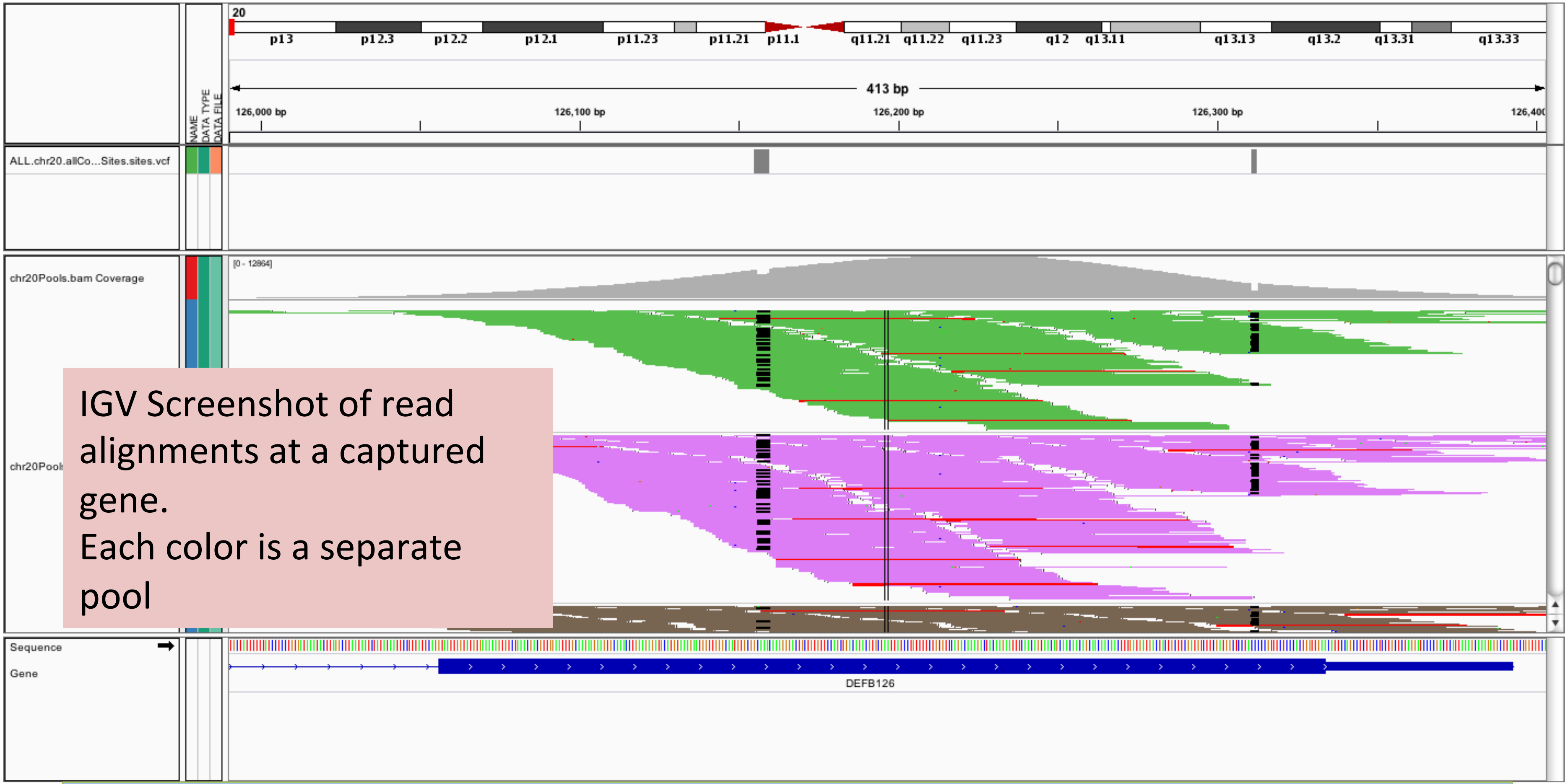
Variable	Result
Successfully sequenced pools	90/92 (97.8%)
Data available now	24 HiSeq lanes (2x76bp reads) for pooled samples. 2 HiSeq lanes (2x76bp reads) for individual barcoded samples
Percentage of reads aligned	98.8 %
Mean Insert size	176
% of total reads from NA12878	19 %
Initial target size	50133
Targets uniquely mapping to genome and with some data	48751

We achieved high depth of coverage in most designed baits



Mean DOC across ALL ~48K baits is about 65,000x across all pools and reference sample (~700x per pool).

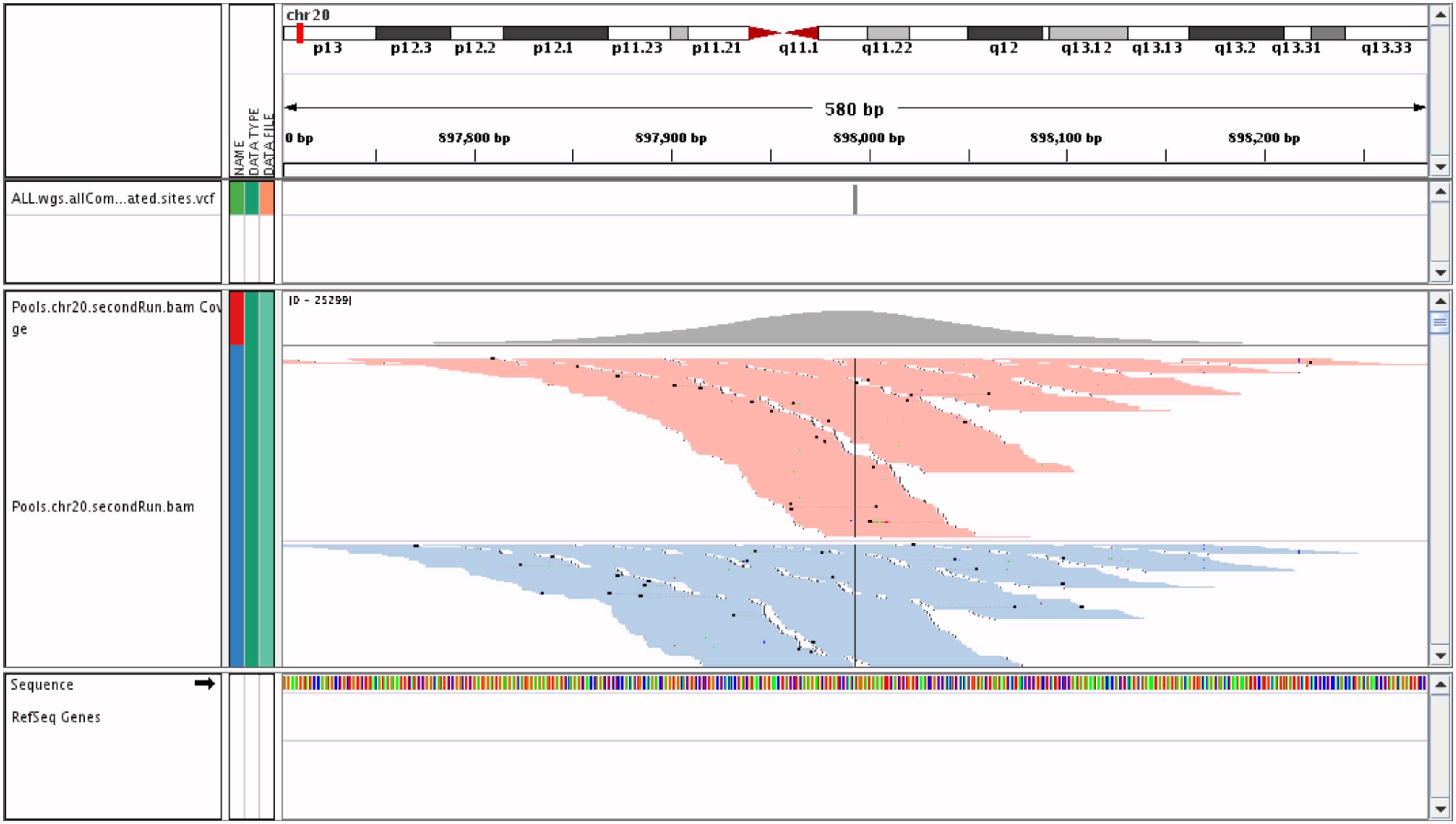
Resulting reads show successful capture and sequencing around targeted variant sites



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

A false positive in 1000 Genomes



A SNP from 1000G with no apparent support in reads

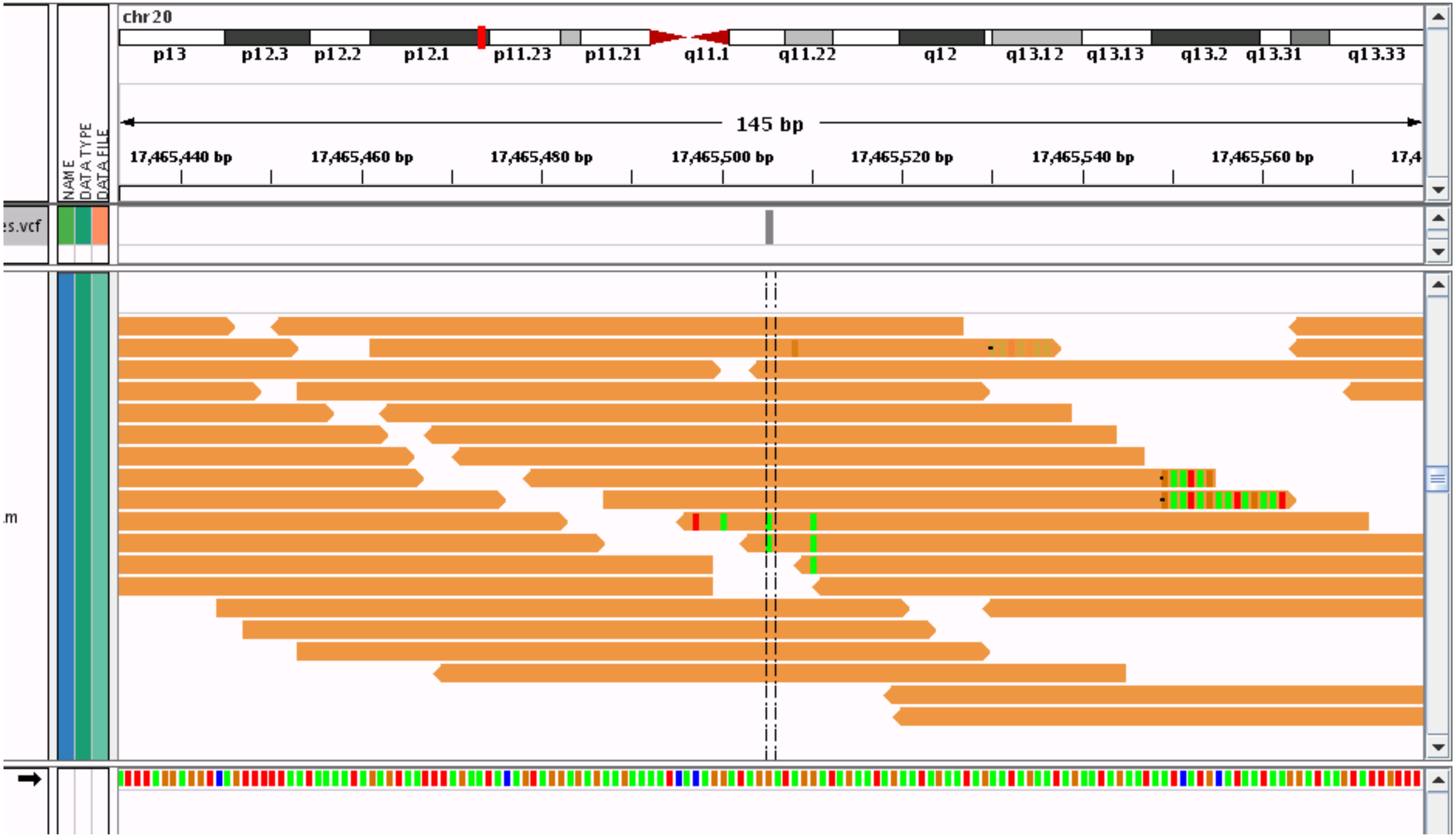
Calls at control sites show that we can discrimate true and false variation

Set	Pool Caller called Monomorphic (AC=0)	Pool Caller called Polymorphic (AC>0)	No-call/Filtered (not enough coverage)
OMNI Mono (SNPs)	711	162	109
OMNI Poly (SNPs)	6	956	38
Exome Chip (SNPs)	3	956	41
Mills Indel Chip	14	940	46

Notes:

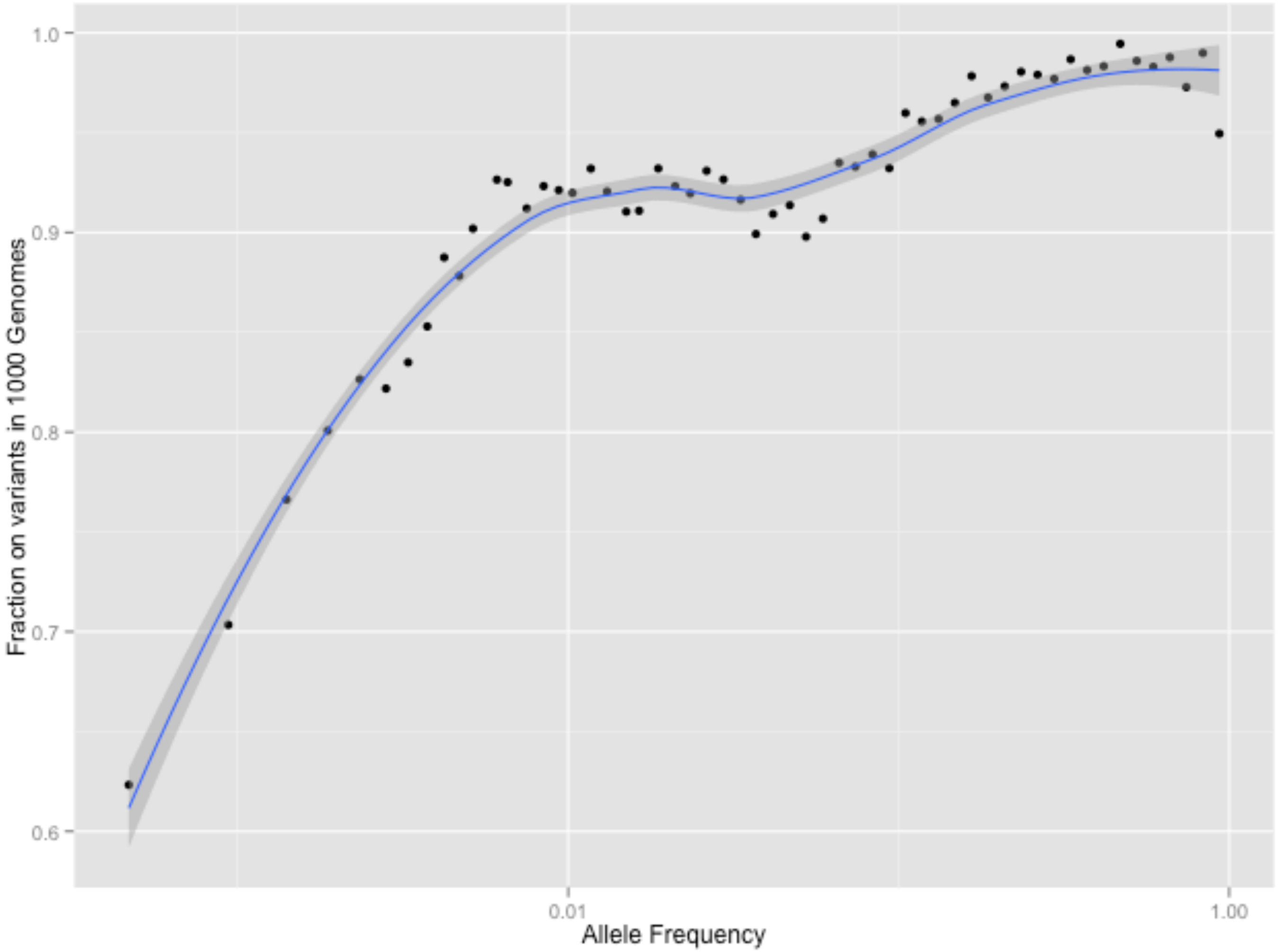
- 3 Exome Chip SNP sites called monomorphic shows that caller is doing what it's expected to do: no evidence of polymorphism in 1000G samples.
- OMNI monomorphic sites which were called polymorphic:
  - Hard to call sites that are ambiguous and possibly should have been filtered out (~120 sites).
  - Sites where there's clear variation but called SNP is wrong allele (~40 sites).

An OMNI monomorphic site that we called polymorphic is a hidden large indel



Single "SNP" is in about 10% of reads. HaplotypeCaller discovered 15 bp insertion at site! Suggests a clear future direction of integrating HaplotypeCaller into framework.

Well over 90 % of all SNPs called by Pool Caller with AF > 1% are already in 1000 Genomes



85,159 SNPs called in all designed baits and filtered by standard VQSR and depth

1000 Genomes SNP and Indel site validation consistent with published rates

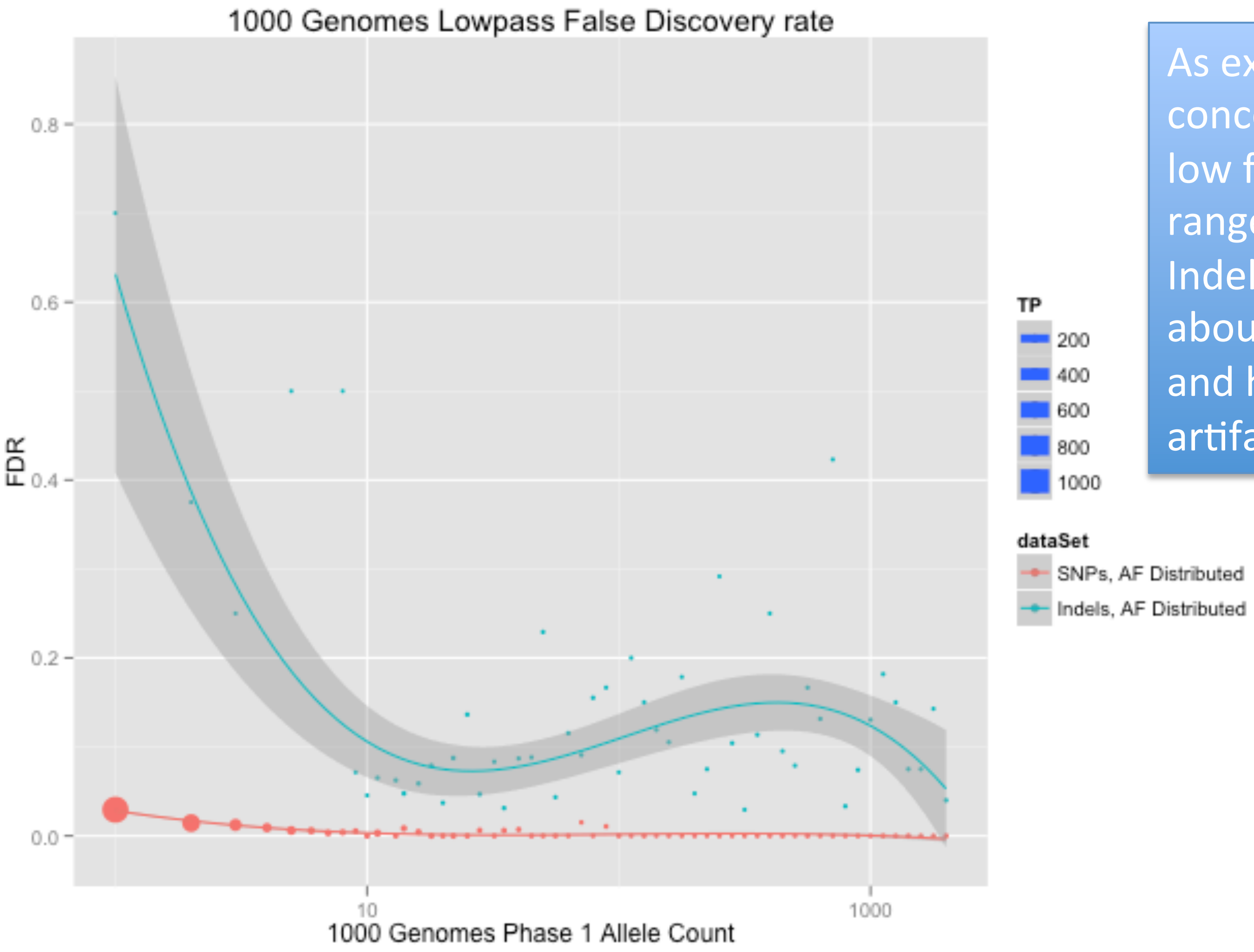
Data Set	# of called sites [1]	Lenient FDR (%)
AF SNPs	6166	1.9 %
Uniform SNPs [2]	5963	5.5 %
AF Indels, post-SVM filtering	1326	18.0 %
AF Indels, pre-SVM filtering [3]	3591	39.2 %
Uniform Indels, post-SVM filtering [2]	2192	17.8 %
Uniform Indels, pre-SVM filtering [3]	3131	36.5 %

NOTES:

1. Only validation sites that had total depth > 5000 and Reference Sample Depth > 500 were kept. **Total yield of about 70-75 % of initial validation targets**
2. One of the validation samples was found later on to have systematic sequencing issues, so the uniformly picked sets **may** have higher SNP error rate due to this.
3. Bait design and site selection were done after preliminary V3 integration was done, but before final SVM filtering removed many indels.

Low-pass FDR in Nature 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

Other sites included for validation are interesting as well.

Data Set	# of called sites (1)	FDR (lenient) (%)
LOF SNPs	5207	5.7 %
LOF Indels	7760	63.2 %
LOF SNPs polymorphic in Phase 1 release (2)	5185	5.6 %
LOF Indels polymorphic in Phase 1 release (2)	989	22.5 %

NOTES:

1. Only validation sites that had total depth > 5000 and Reference Sample Depth > 500 were kept.
2. Many LOF Indels didn't get to be in final Phase 1 integrated set since exome-only indels weren't integrated.

Future Work and extensions

- Application to clinical problems at a large scale.
- Detailed analysis:
  - 1000 Genomes validation rate by event size/functional type, etc.
  - Strict vs. lenient allele matching.
  - Investigation of error modes ("why did we call each particular FP?")
  - Large deletion analysis
  - Concordance with Illumina Exome Chip indels?
- Methods optimization:
  - Use of new GATK local-assembly based approach jointly with new analytics.
  - Better estimation of site error models.

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!**