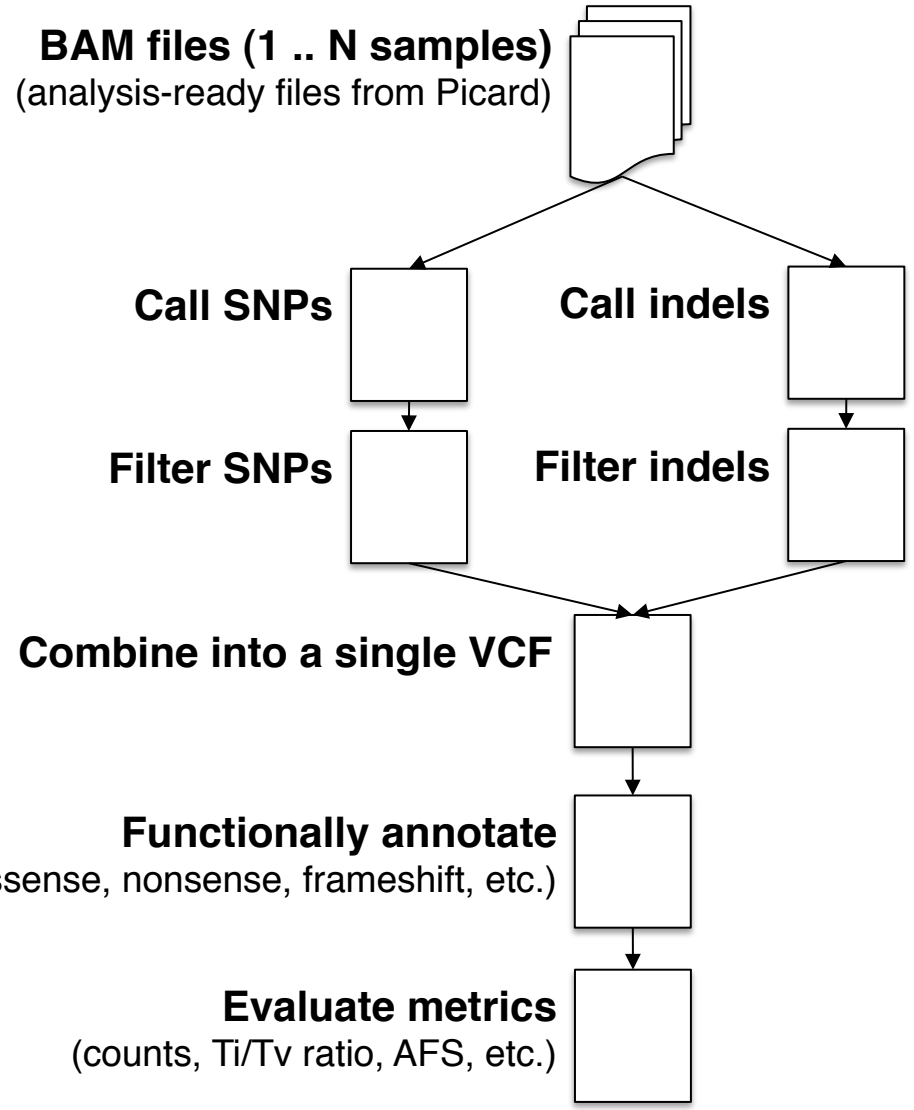
MPG Workshop / February 17, 2011

QPipeline: variant calling pipeline for custom-target, exome, and deep whole-genomes

Kiran V Garimella (kiran@broadinstitute.org)
GENOME SEQUENCING AND ANALYSIS, BROAD INSTITUTE

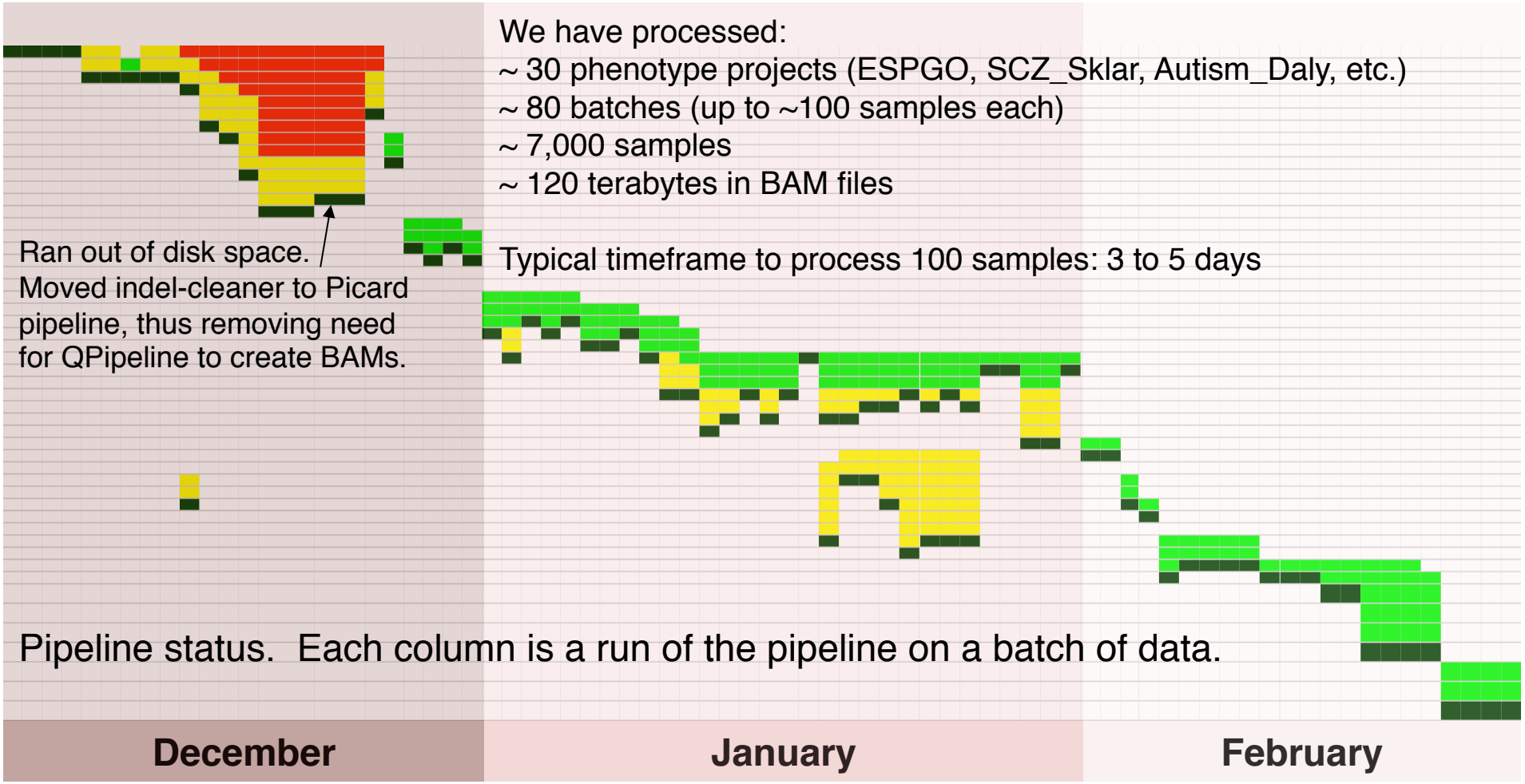
GATK tools enable discovery of interesting variation; QPipeline applies them to nearly all MPG NGS projects

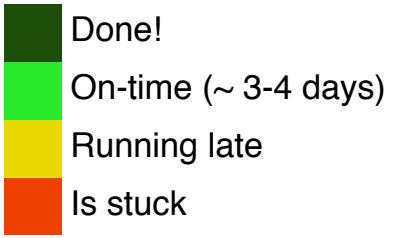


QPipeline:

- Takes Picard's analysis-ready BAM files
- Runs best-practice SNP and indel-calling protocol (multi-sample variant calling and genotyping, latest filters)
- Annotates functional consequences for SNPs (missense, nonsense, silent) and indels (frameshift, inframe, noncoding) for ***all*** transcripts
- Runs on custom-target, whole-exome, and deep whole-genome data
- Is fast (~ 3-5 days/100 samples)

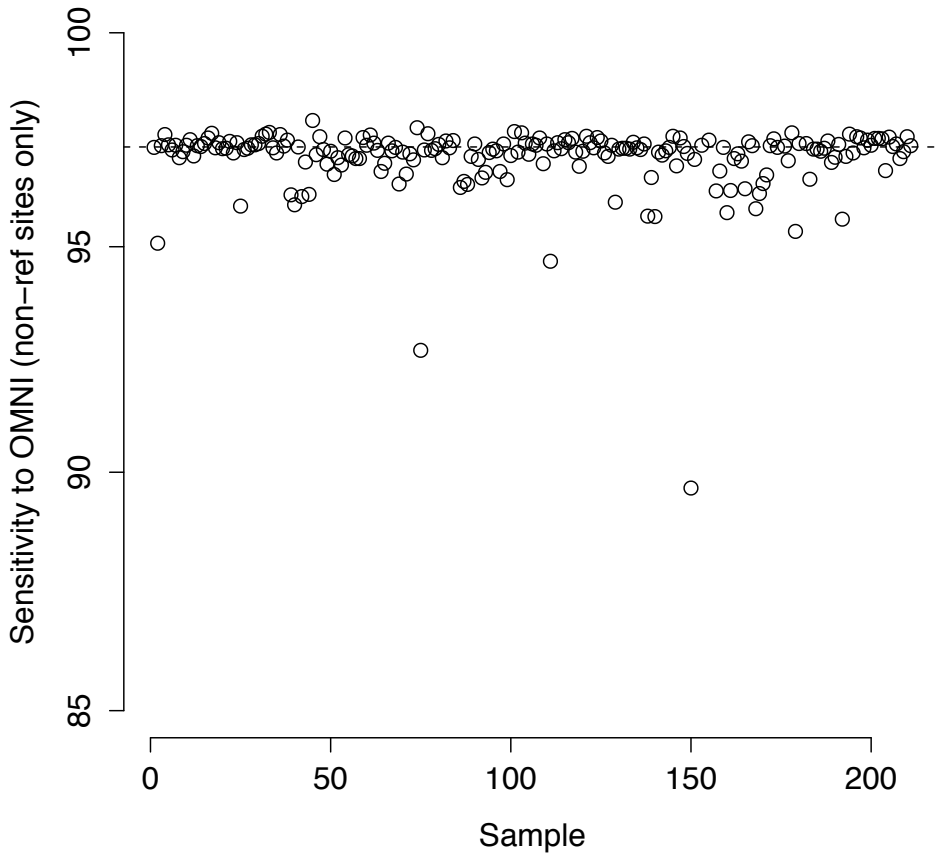
What we've processed: a 75-day history of the pipeline

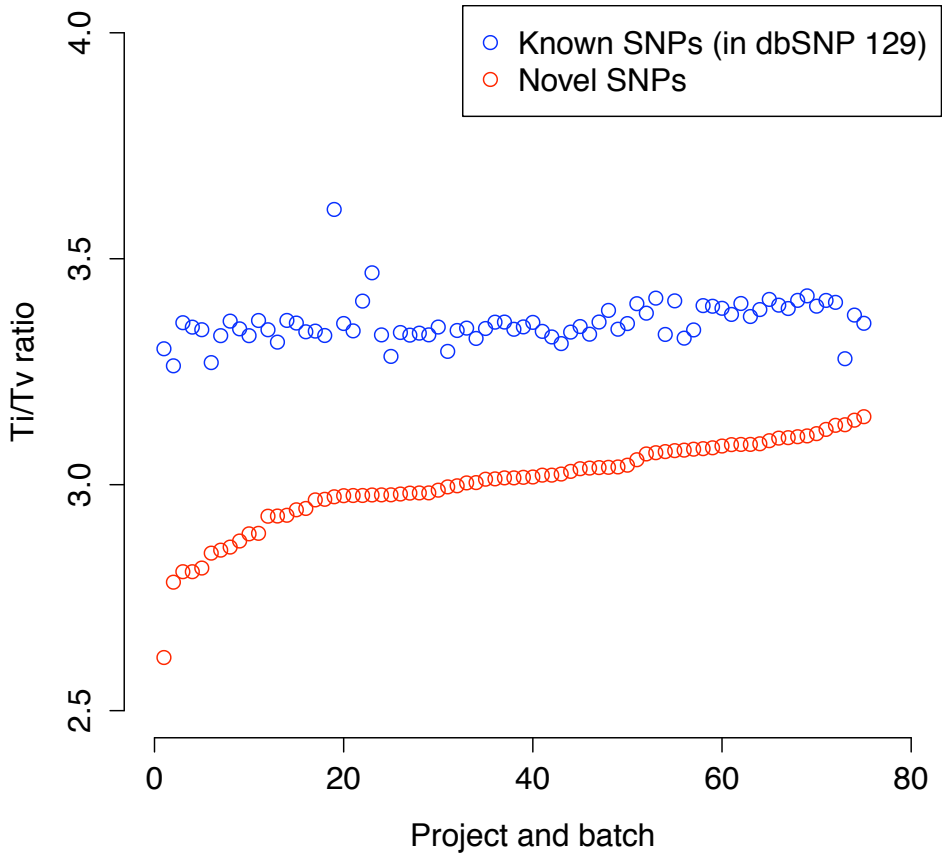




Pipeline status is updated every workday by 11am. Available as a Google document at https://docs.google.com/a/broadinstitute.org/.

High sensitivity (~97%), low FPR (5-9%), and consistent performance across projects and batches

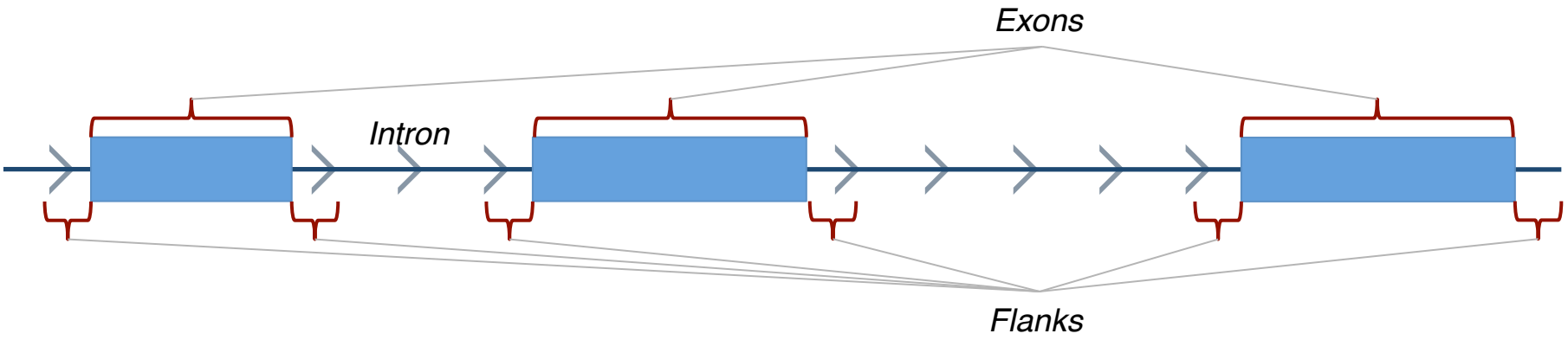




Using exomes from the 1,000 Genomes Project and comparing results to the OMNI chip (~55k variants, many of which are low-frequency), variant sensitivity is ~97-98%. Estimated false-positive rate (from Ti/Tv) is less than 10%.

Starting Friday (2/18/2011), all MPG projects will be reprocessed to include calls in exons and 50-bp flanks

	Region	Expectation[1]	All		Known		Novel	
		Ti/Tv	SNPs	Ti/Tv	SNPs	Ti/Tv	SNPs	Ti/Tv
Exons only	32,950,014 bp	~3.2	209,709	3.18	80,135	3.38	129,574	3.06
Flanks only	19,200,667 bp	~2.3	152,051	2.36	51,482	2.40	100,569	2.34
Exons and flanks	51,766,768 bp	~2.8	360,904	2.79	131,262	2.93	229,642	2.71



Expanding by 50-bp on either side of the exons allows us to recover some of that variation, allowing us to capture variants in splice regions.

1. Expected Ti/Tv based on the 1KG Phase 1 August 2010 whole-genome release (BC + BI + NCBI + UM 2-of-4 way consensus)

Contact us to get data into pipeline

Contact our data manager, Corin Boyko (corin@broadinstitute.org), to set up a 15-minute sample upload and QC appointment

- By setting up the project definition together, we:
  - Can go over initial QC metrics together and discuss whether outlier samples should be dropped
  - Can ensure that batch definitions have no overlapping samples
  - Can ensure that batches are right size (currently ~100 samples/batch)
  - Can discuss whether the absence of fingerprints will be a problem (generally not an issue if GWAS data is available)
  - Can avoid a long email chain wherein we can get confused as to what needs to be done

- Each project will get a unique identifier (e.g. PXQ74, PY2M8). All pipeline output for the project will be available under our pipeline root directory, e.g.:

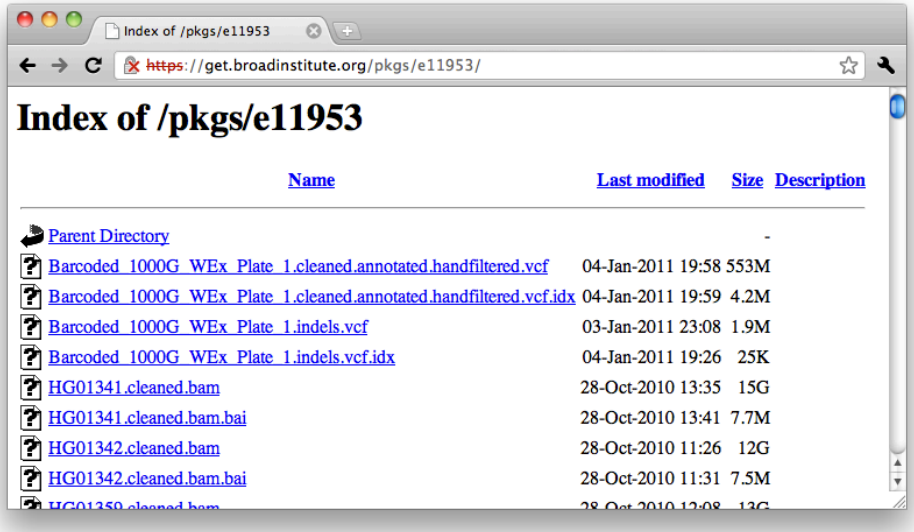
```
/humgen/gsa-pipeline/PXQ74
/humgen/gsa-pipeline/PY2M8
```

Three steps to set up a data transfer to colleagues outside the Broad

1. Create a file manifest - two column file with path and name, e.g.

```
/path/to/my.file.bam              my.file.bam
/path/to/my.file.bai              my.file.bai
/path/to/NA12878.vcf              NA12878.vcf
/path/to/different/NA12878.vcf    NA12878.other.vcf
```

2. Send manifest to Help@Broad, cc'ing PM and PI.
   Help will set up the secure-HTTP server with links to the files listed in the manifest and send back
   username and password.



3. Credentials and web address
   can then be sent to colleagues.
   All files can be downloaded by
   a single Unix wget command, i.e.

```
$ wget --no-check-certificate --mirror \
       --user e11953 --password 1C0ni9qWJOMe \
       https://get.broadinstitute.org/pkgs/e11953
```

Conclusion

- QPipeline produces best-practice SNP and indel calls in a matter of days
  - Works for custom target, whole-exome, and deep whole-genome
  - Consistent results across projects and batches

- All projects receive updates when pipeline changes significantly
  - New indel caller will be back-propagated to all projects
  - Variant calls will be made in 50-bp-expanded target regions

- Easy to get data into and out of pipeline
  - To initiate pipeline on your project, set up a meeting with us and we'll enter samples and go over metrics together
  - To transfer data to colleagues outside the Broad network, create a two-column manifest (file name, display name) of the files you want to transfer and send it to Help@Broad.

Acknowledgements

IT/Systems

Aaron Ball
Eric Jones
Matthew Trunnell

Production Informatics

Kathleen Tibbetts
Alec Wysocker
Seva Kashn
Zach Leber
Tim Fennell
Toby Bloom

Genome Sequencing and Analysis

Corin Boyko
Chris Hartl
Khalid Shakir
Mauricio Carnerio
Matt Hanna
Ryan Poplin
Guillermo del Angel
Menachem Fromer
Eric Banks
Mark DePristo

All the pipeline users who have given us feedback, including

Ben Neale
Ron Do
James Pirruccello
Sarah Calvo
Shaun Purcell

Appendix

Transition/transversion ratio (Ti/Tv) can be used to estimate false-positive rate

- The observed Ti/Tv is the result of a mixture of true-positive SNPs and false-positive SNPs.

- Assume that the Ti/Tv that should be observed if one only includes true-positives is that of the known SNPs in the target region (for whole-exome data, $Ti/Tv_{TP} = 3.3$).

- Assume that the Ti/Tv that should be observed if one only includes false-positives is that of a purely random event (for each base, 1 transition mutation and 2 transversions possible, hence $Ti/Tv_{FP} = 0.5$).

- Thus,

$$\alpha(Ti/Tv_{TP}) + (1-\alpha)(Ti/Tv_{FP}) = Ti/Tv_{observed},$$

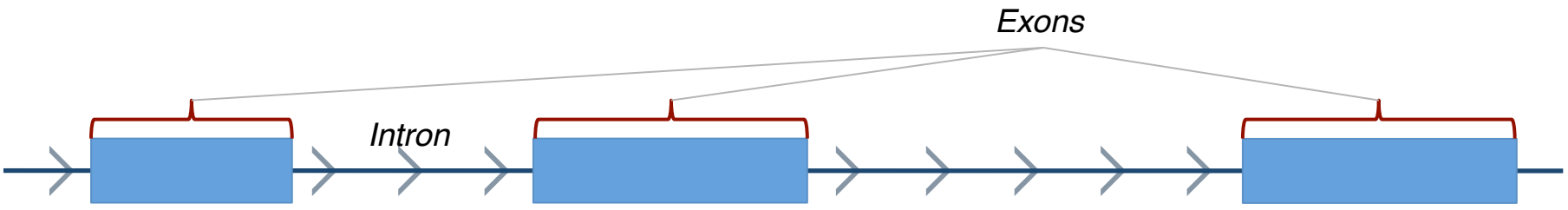
  where α is the mixture coefficient specifying the proportion of SNPs that are true-positives.

- The term $(1-\alpha)$ specifies the proportion of false-positives, so

$$(1-\alpha) = \%FP = 1 - (Ti/Tv_{observed} - Ti/Tv_{FP})/(Ti/Tv_{TP} - Ti/Tv_{FP})$$

Previously, variants were only called in the explicitly specified targets (exons + 2-bp flanking sequence)

	Region	Expectation[1]	All		Known		Novel	
		Ti/Tv	SNPs	Ti/Tv	SNPs	Ti/Tv	SNPs	Ti/Tv
Exons only	32,950,014 bp	~3.2	209,709	3.18	80,135	3.38	129,574	3.06



We currently stay within the “safe” confines of the target, but in doing so, we skip some potentially interesting variation in the splicing and promoter regions.

1. Expected Ti/Tv based on the 1KG Phase 1 August 2010 whole-genome release (BC + BI + NCBI + UM 2-of-4 way consensus)