Xavier Estivill & Jan Korbel, cancer germline genome update, on behalf of the PCAWG germline working group (PAWG-8)

Annai Systems – Francisco De La Vega, Ying Wu

Broad Institute – Ayellet Segre, Adam Kiezun, Gad Getz

Catholic Medical Center – Yeun-Jun Chung

CRG Barcelona – Oliver Drechsel, Stephan Ossowski, Georgia Escaramis, Xavier Estivill

DKFZ – Matthias Schlesner, Ivo Buchhalter

EMBL – Sebastian Waszak, Joachim Weischenfeldt, Sergei Iakhnin, Serap Erkek, Jan Korbel

Cambridge University – Jamie Allen, Douglas Easton

Hallym University – Ji Wan Park, SG Luke Heo, Eun Pyo Hong

Kiel University – Reiner Siebert

MD Anderson Cancer Center – Ken Chen, Bo Peng

National University Hospital Korea – Youngil Koh, Sungsoo Yoon, Mi Kyeong Lee

Sanger Institute – Jose Tubio, Young Seok Ju, Erik Garrison

Samsung Medical Center – Youngwook Kim, Keunchil Park

Stanford University – Suyash Shringarpure, Carlos Bustamante

U of Geneva – Olivier Delaneau

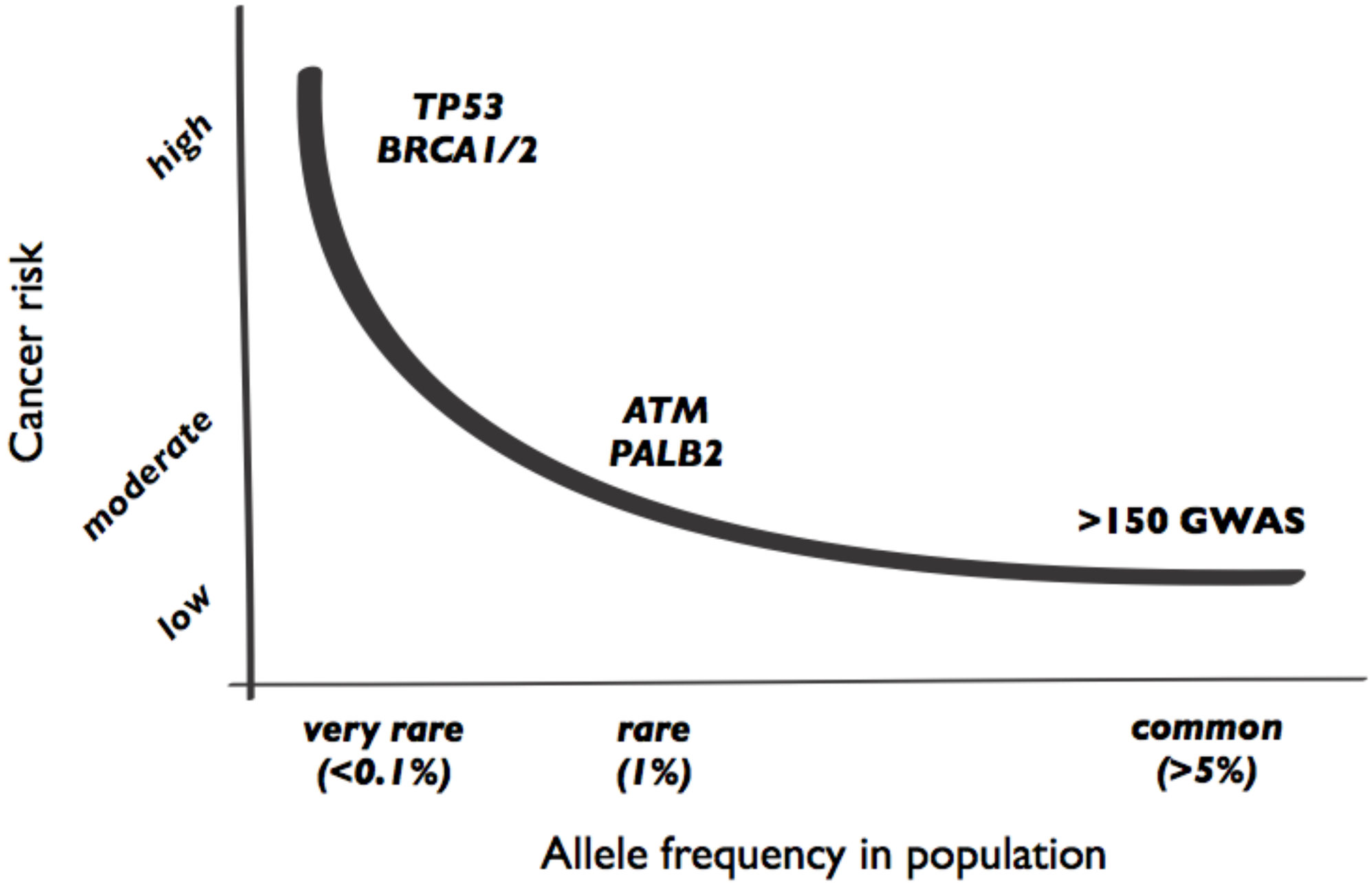
U of Utah – Gabor Marth

WashU – Reyka Jayasinghe, Li Ding

Germline WG co-chairs: Xavier Estivill & Jan Korbel

Verona, 15th February 2015

Genetic predisposition to cancer



Objectives – PCAWG germline genome working group

Study of the "other" PCAWG genomes

Technically-oriented

- Infer hereditary / "germline" polymorphisms from the genomes of matched normal tissue samples (sequenced at ≥30x coverage).

  - Generate high-quality variant callset including SNPs, InDels, structural variants (PCAWG germline variant callset).

  - Haplotype-block phased variants to achieve particular utility for genetics studies (SNP imputation / association studies).

Research questions

- Investigation of cancer germline susceptibility loci.
- Study interactions between germline and somatic genetic variants.

Germline Working Group: Expected Scientific Outputs

(A) Define landscape of germline mutations across cancer types

- Identify cancer risk genes enriched in rare, damaging germline mutations.
- In-depth investigation of susceptibility genes & pathways.

(B) Germline-somatic genome associations across cancer types

- Links between germline variants and somatic mutation & DNA rearrangements patterns (e.g. mutational signatures, global or regional / haplotype-specific mutational effects).
- Links between germline variants & gene expression / DNA methylation (eQTL/meQTL mapping, allele-specific analyses).

Current status: analysis workflow implemented based on pilot-63 dataset

Raw germline variant discovery (PCAWG pipelines)

VCF VCF VCF VCF VCF VCF VCF VCF VCF VCF

Variant integration (ensemble allele list)

Ensemble allele annotation in each sample & genotyping

VCF VCF VCF VCF VCF VCF VCF VCF VCF VCF

Machine learning (SVM)-based variant filtering*

VCF VCF VCF VCF VCF VCF VCF VCF VCF VCF

Variant haplotype-phasing (informed by tumor type)

Final VCF

*Unless filters defined / set through calling algorithm

<u>Germline variant callers tested for PCAWG pilot-63 set:</u>

(respective germline VCFs available at NCI jamboree):

HaplotypeCaller/Broad (SNP)

Platypus/DKFZ-EMBL (SNP/InDel)

Samtools/DKFZ-EMBL (SNP)

rtg/Annai Systems (N=1452 complete) (SNP/InDel)

VarScan/WashU (SNP)

Clindel/CRG (InDel)

Delly/EMBL (SV)

CNVnator/Yale (SV)

PeSVfisher/CRG (SV)

Pindel/Sanger/WashU (InDel/SV)

BreakDancer/WashU (SV)

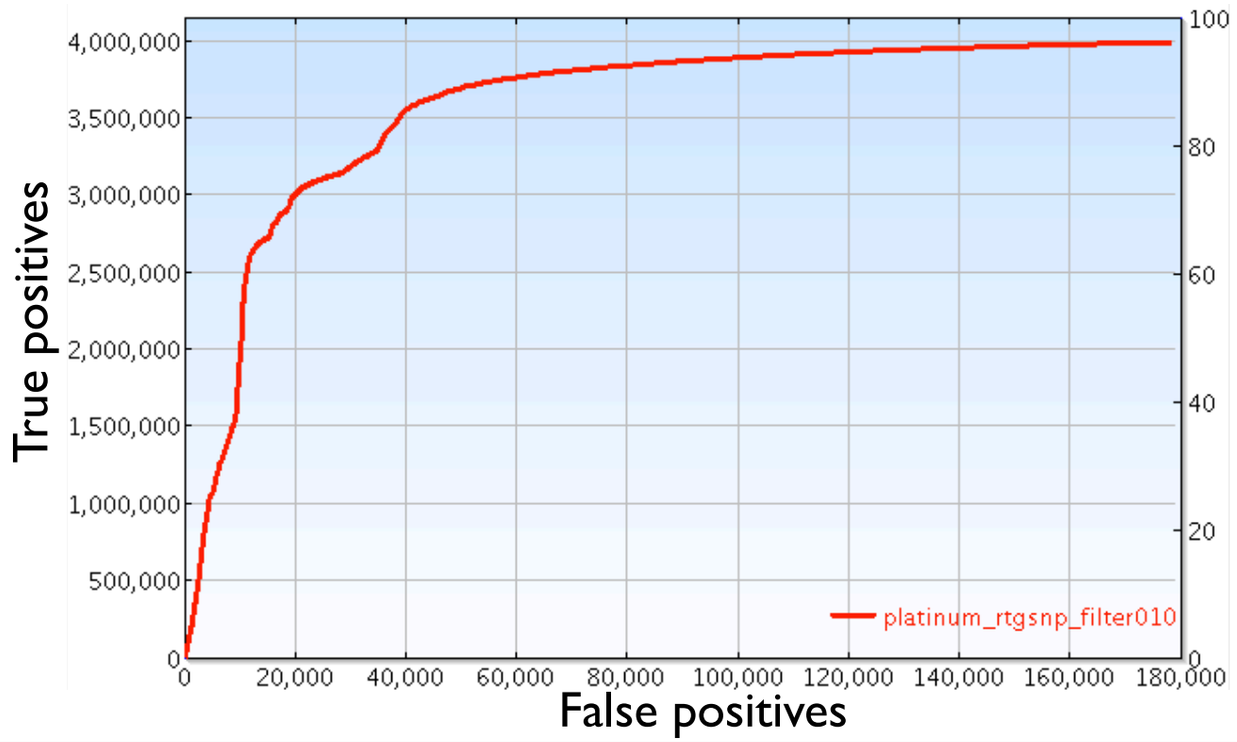
Sebastian Waszak, Erik Garrison

Germline variant validation approaches

1. Generate validation data for *N*=3000 variant sites at WashU (picking of variant sites is variant allele frequency weighed).
2. Deploy our pipeline on external reference sets:
   - Mendelian concordance tests in "Illumina platinum genome" and 1000 Genomes Project samples (*N*=3 parent-offspring trios).
   - Compare with *Genome in a Bottle* truth data.
3. Measure concordance with SNP genotype arrays available for subset.
4. Assess variant haplotype-phase accuracy through imputation.

Example: ROC curve – based on Genome-in-a-bottle data for individual with European ancestry (NA12878)

RTG germline SNP caller



SVM-based filtering of raw germline calls

1. DKFZ raw germline callset based on pilot63 samples

2. Site-level annotation using *freebayes* and based on a 26-feature vector (for example: total read depth, read mapping quality)

3. SNP/MNV filtering using probabilistic support vector machine (SVM):

    1. PCAWG alignment and variant calling pipeline run on external reference genome of European ancestry (NA12878, 1000 Genomes Project/ Platinum Genomes)

    2. Ground truth (TP/FP SNV-MNVs) based on NA12878 high confidence callset (NIST-GIAB + RTG/Illumina Platinum Genomes)*

    3. SVM trained/tested on a random subset of 10,000/10,000 sites

*ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/variant_calls/GIAB_integration/

Sebastian Waszak, Erik Garrison

Initial results for germline SNPs

Comparison of representative pilot germline SNV/MNV callsets (chr20 of pilot-63)

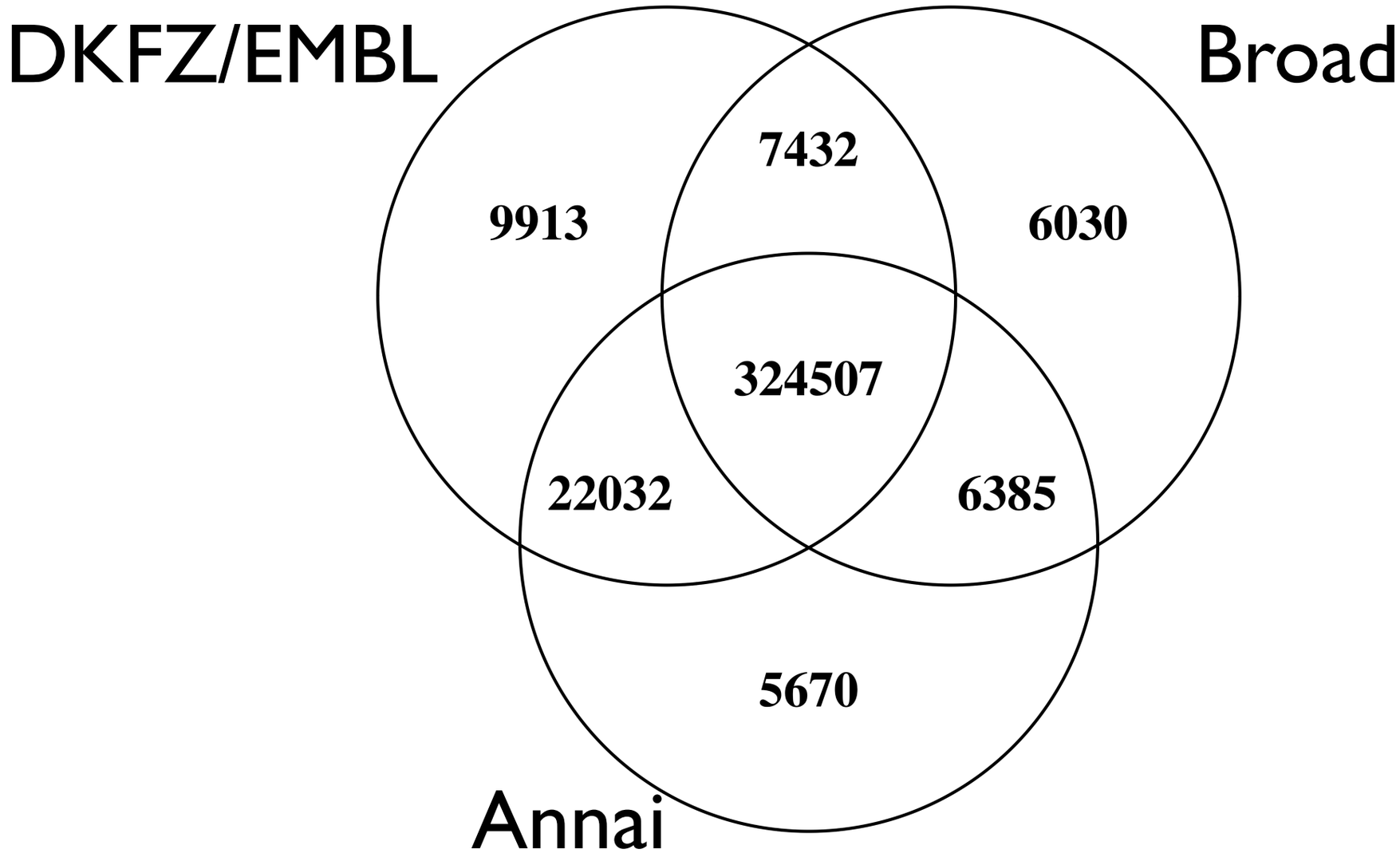
Center	Total # alleles*	Known alleles	Novel alleles	ts/tv*	# genes affected by alternate alleles
Broad	344,354	299,840 (87.1%)	44,514 (12.9%)	2.27	1,024
Annai Systems	358,594	311,254 (86.8%)	47,340 (13.2%)	2.30	1,025
DKFZ/ EMBL	363,884	312,784 (86.0%)	51,100 (14.0%)	2.30	1,025

*Multi-allelic records broke into multiple records and complex variants decomposed into canonical alleles / variant annotation based on ENSEMBL release 78

*expected ts/tv ratio for SNPs on chr20 is ~2.3

Sebastian Waszak

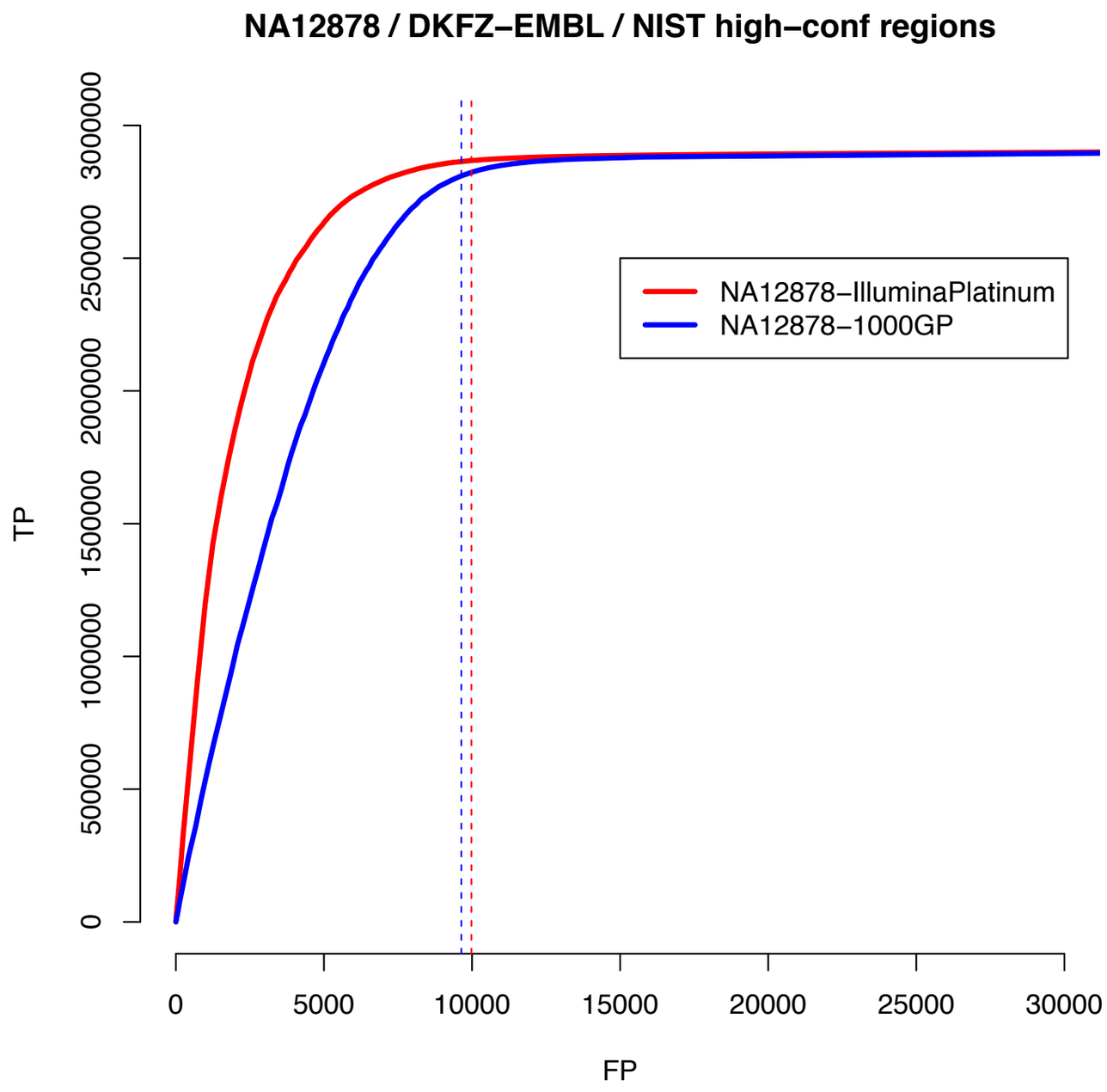
Overlap between representative pilot germline SNV callsets (chr20 of pilot-63)



94.3% of germline SNV alleles have been called by two or more pipelines

Sebastian Waszak

SVM filter applied on two independent raw DKFZ-SNV/MNV callsets for NA12878



Training:

- NA12878-IlluminaPlatinumGenomes
- 10,000 random sites
- NIST-defined high-confidence regions

Testing:

- NA12878-IlluminaPlatinumGenomes
- NA12878-1000GP
- All sites called in high-confidence regions

SVM cutoff > 0.1:

- #false positive calls
  - Illumina: 9,980
  - 1000GP: 9,638
- #true positive calls
  - Illumina: 2,867,633
  - 1000GP: 2,810,067

SVM cutoff > 0.1:

- FDR
  - Illumina: 0.35%
  - 1000GP: 0.34%
- Sensitivity
  - Illumina: 96.60%
  - 1000GP: 94.66%

Sebastian Waszak, Erik Garrison

Discussion / Reminders for groups calling germline variants

1. Mendelian concordance in *N*=3 Illumina “platinum genome” parent-offspring trios aligned at CRG.
   Please do run your germline pipeline on these samples too!

2. Inclusion of germline SNP6 array data for verifying SNP genotype concordance in PCAWG dataset?
   (These are available for approximately half of all TCGA samples, and further for several ICGC samples (*e.g.* a subset of medulloblastomas).)

3. Interest in inclusion of larger panel of cancer exomes, including ICGC and TCGA exomes - plan: to be analyzed with local resources at CRG.

4. Interest in WXS variants from Exome Aggregation Consortium (ExAC) with over 60,000 exomes (including many non-cancer samples), as controls from several ethnics groups are needed.

--> Lack of non-cancer non-cancer WGS sample sets is a main barrier in the investigation of cancer germline susceptibility loci

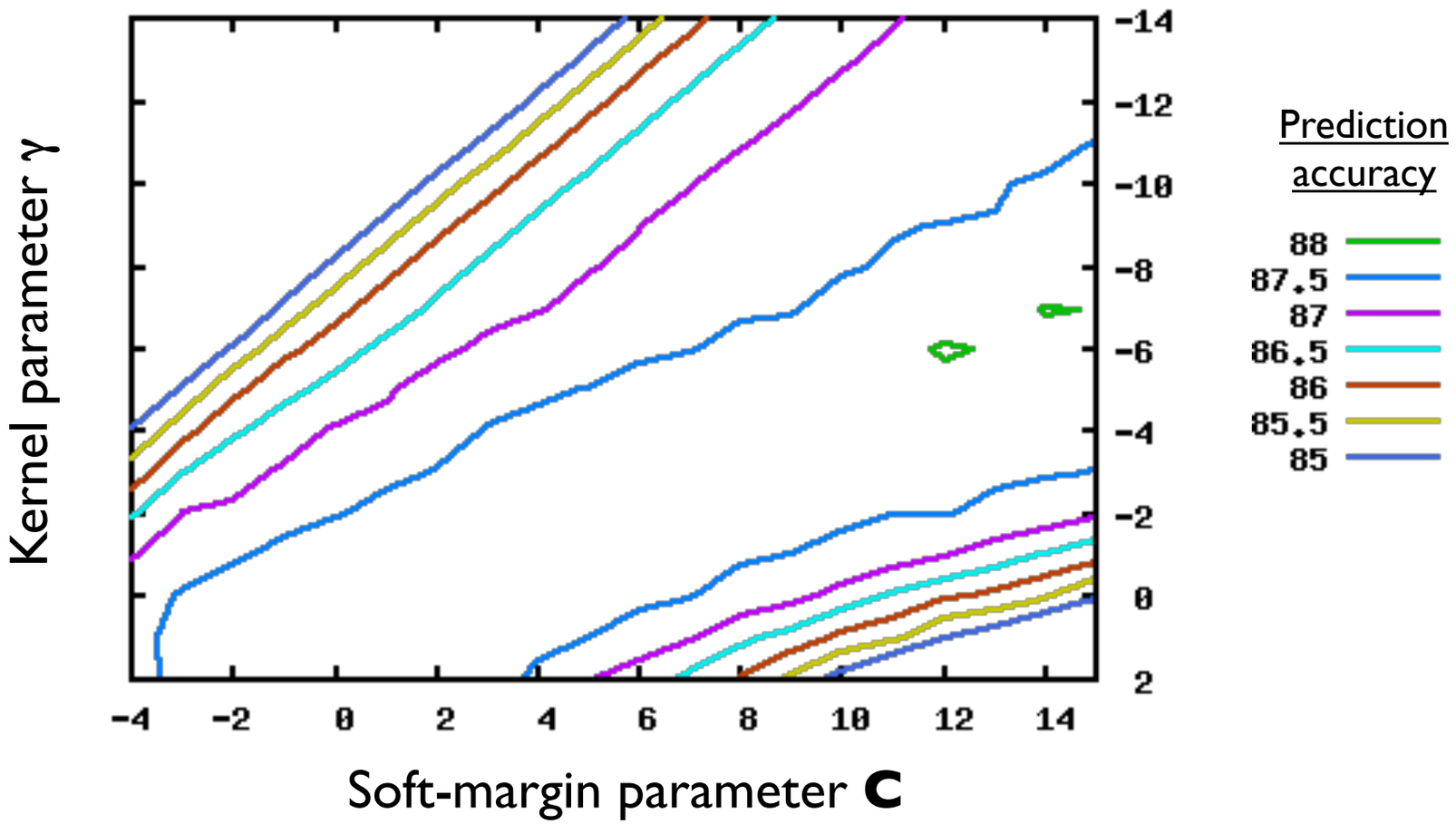
BACKUP SLIDES

Site-feature vector for SVM-based variant classification

- **DP**: Total read depth at the locus
- **RO/AO**: Ref/alt allele observation count
- **QR/QA**: Ref/alt allele quality sum in phred
- **SRF/SRR/SAF/SAR**: Number of ref/alt observations on the fwd/rev strand
- **SRP/SAP**: Strand balance probability for the ref/alt allele
- **AB/ABP**: Allele balance (probability) at heterozygous sites
- **RPL/RPR**: Read placed left/right
- **RPP/RPPR**: Read placement probability (for ref observations)
- **EPP/EPPR**: End placement probability (for ref observations)
- **ODDS**: The log odds ratio of the best genotype combination to the second-best
- **MQM/MQMR**: Mean mapping quality of observed ref/alt alleles
- **PAIRED/PAIREDR**: Proportion of observed ref/alt alleles which are supported by properly paired read fragments
- **EntropyCenter**: Entropy of centered sequence of 10bp
- **QUAL**: quality score

Sebastian Waszak, Erik Garrison

SVM parameterization for DKFZ SNV/MNV calls



Grid search with 5-fold cross-validation to identify best SVM parameterization for DKFZ SNV/MNV calls

Sebastian Waszak, Erik Garrison