Services Presentation

INTEGRAGEN

Genomics Experts

Illumina Seminar – Marriott – May 11th

IntegraGen at a glance

INTEGRAGEN

Autism

Oncology

Genomics Services

Serves the researcher's most complex needs in genomics

The n°1 privately-owned genomics platform in France

A Genopole-biocampus company

IntegraGen Services Offering

- High Throughput Genotyping Platform
  - Illumina Genotyping Platform
  - Other Material
  - Bioinformatics & Biostatistics experts in association studies

IntegraGen has the capacity for running and analyzing any kind of genotyping study

- New Generation Sequencing
  - Illumina GA IIx since March 2009
  - Bioinformatics analysis

For any application

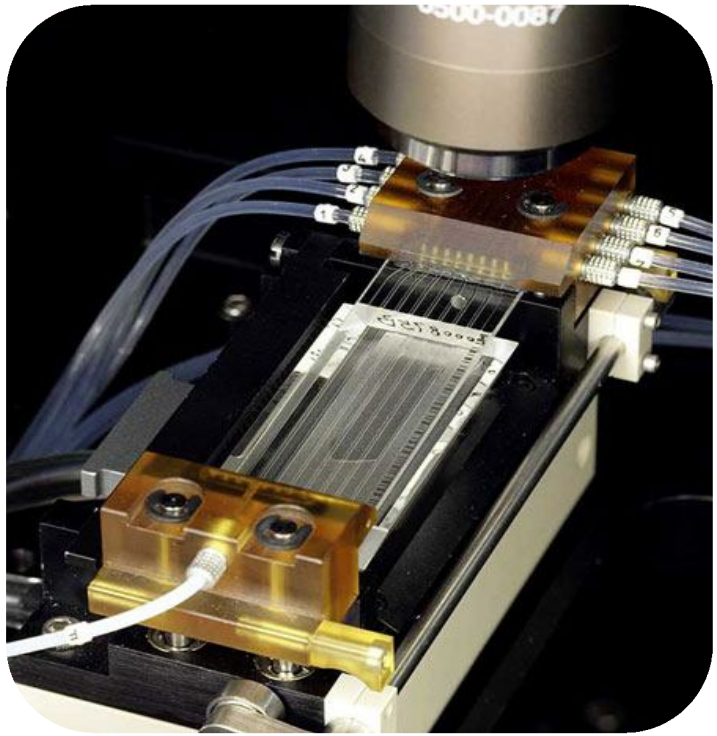
Our Customers

- References 2009
  - La Ligue National contre le Cancer : partner in the CIT program (Carte d'Identité des Tumeurs®)
  - Institut curie
  - Laboratoires Servier
  - INSERM including a large Pharmacogénétique program (iselect)
  - Institut Pasteur de Lille
  - Sanofi-Aventis
  - France Limousine Sélection
  - Limagrain
  - Hospitals…

Scalable Performance

Current Install Specifications (early 2009)

- >50 million reads per flowcell (single read)
- >1.5GB per single read flowcell (36bp read)
- >3.0GB per PE flowcell (36 bp read)
- >750MB/day
- 2 day single read run, 4 day PE run*
- Supported read length: 36bp
- System enabled for 50bp+ reads
- Short Insert Paired End released
- Long Insert Mate Pairs in development

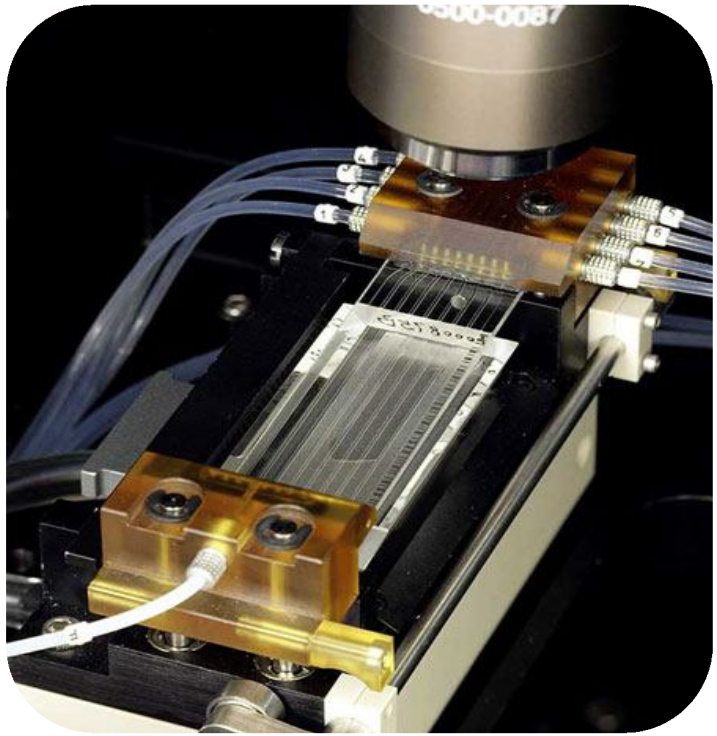


** Short recipe protocol, in final testing currently*

Scalable Performance

Current Install Specifications

- >**200** million reads per flowcell (single read)
- >**20** GB per single read flowcell (**100** bp read)
- >**40** GB per PE flowcell (**100** bp read)
- >**4.5 GB/day**
- ~~2 day single read run, 4 day PE run~~
- Supported read length: **100** bp
- System enabled for **150** bp+ reads
- Short Insert Paired End released
- Long Insert Mate Pairs **ready**

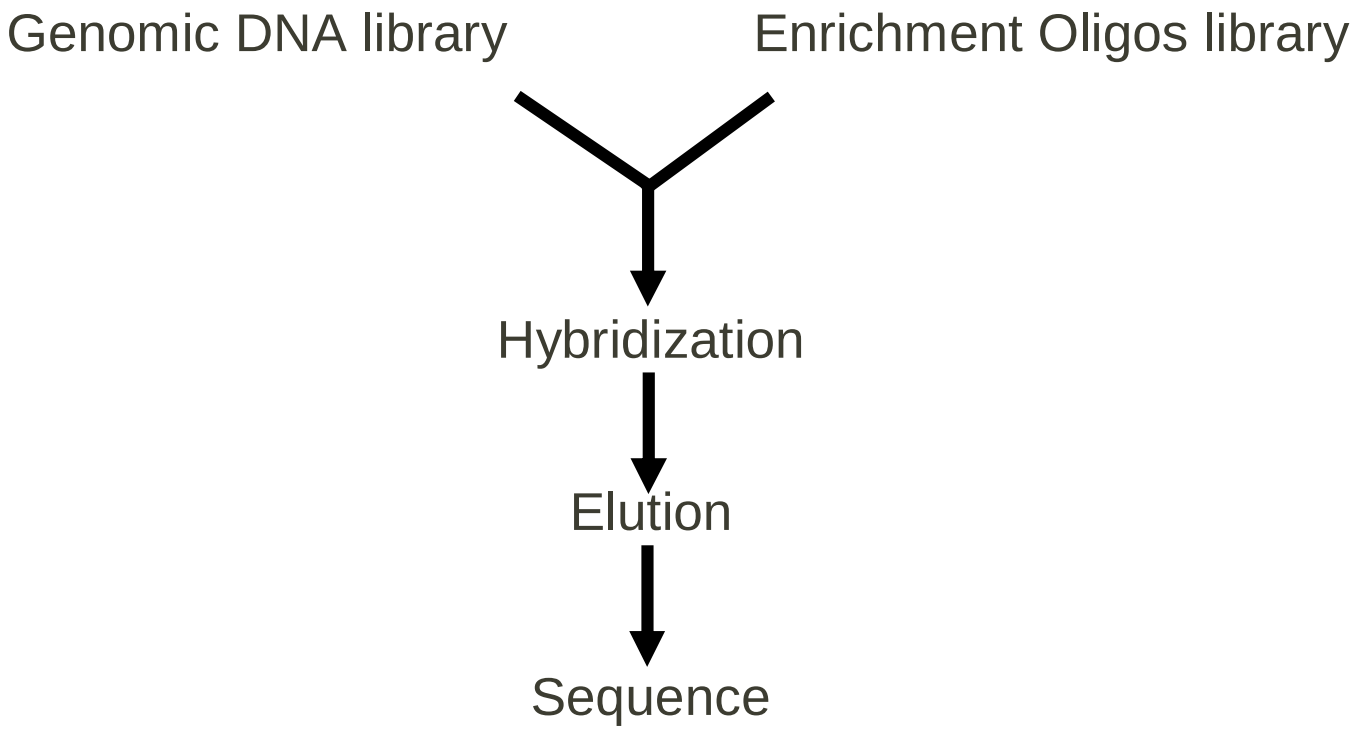


Today's topics

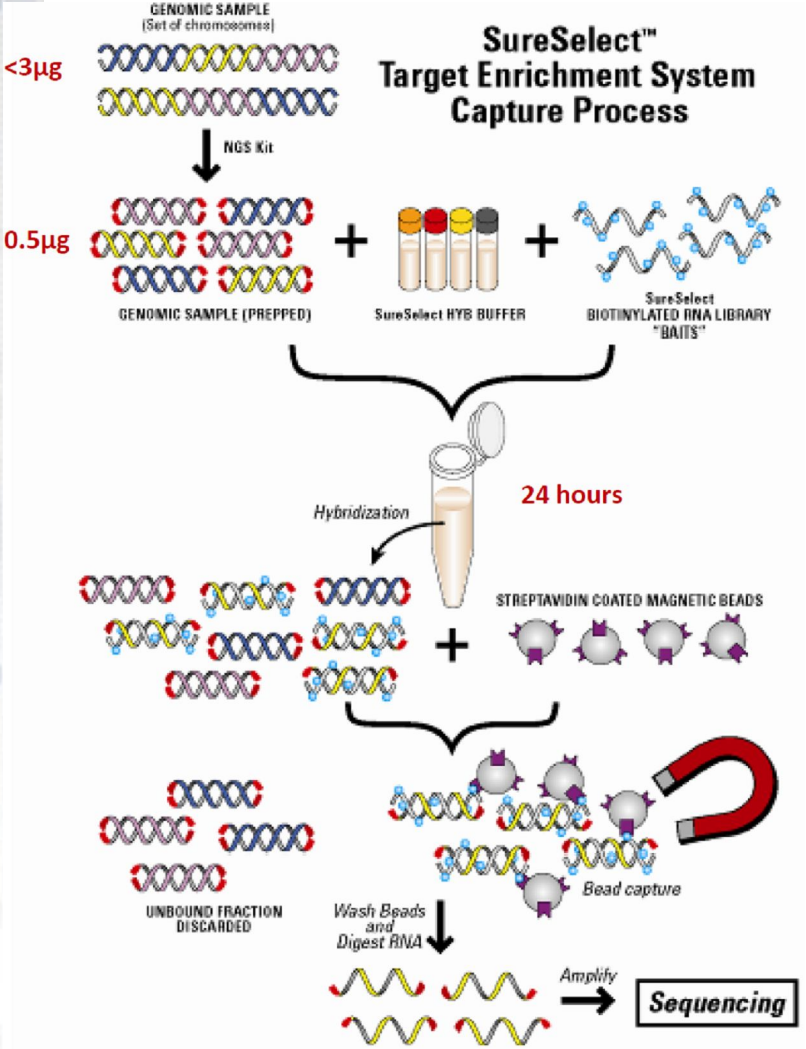
- ReSequencing
  - Enrichment, Capture, Exome, Analysis…

- Rearrangements analysis of tumor cells
- ChIP-Seq
- mRNA-Seq
- DGE
- miRNA

Enrichment by Sequence Capture



Agilent SureSelect enrichment system



Resequencing of 3Mb genomic region identified by linkage study

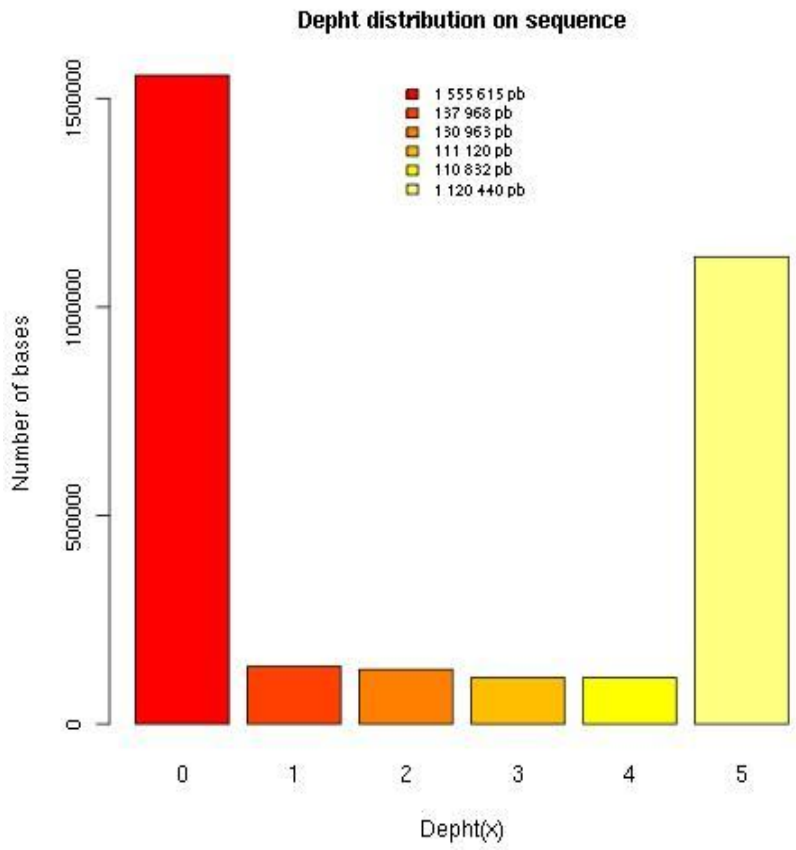
(Pr Sanlaville & Edery – Lyon)

Workflow

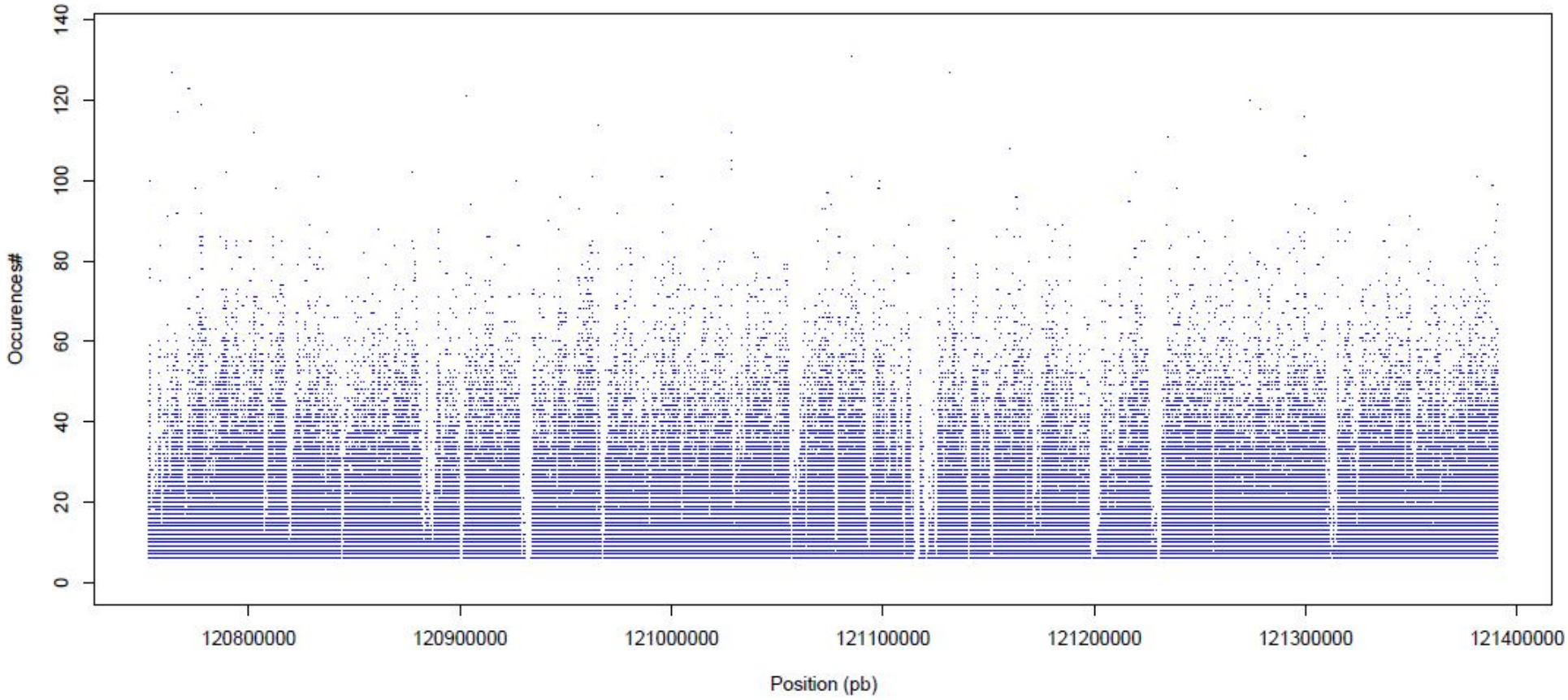
- Submission of the regions of interest through agilent « e-array » web portal. Apply repeat masker.
- Design maximum : 57,750 RNA oligos of 120 bases ~6 Mb,
- 2x to 5x capture depth, corresponding to ~ 3 Mb
- Agilent synthesizes and ships Baits pool of biotinylated RNA (4 – 6 weeks)
- Hybridization against genomic library

Design Results

- Target region: 3,166 Mb
- Design :
  - Baits: 120bases
  - Baits coverage: 5X
  - Baits overlap: 24 bases
  - Centered Design
  - % Region Covered: 50.81%
  - Effective regionsize: 1,609 Mb
- Masked regions: 49.19%
- Sequence:
  - Fragment size: 400 bases
  - 1 sample by Flow-Cell lane
  - Single Read 75 bases



Sequence coverage
represented by start positions



Sequencing QC1
How does the enrichment work?

- Sequence alignment by ELAND (32 bases, 2 mismatchs max)

Samples	Status	# PF clusters	% on Target Region (3Mb)	% on human genome	Specificity (%)	Depth (X)
F01-1	Healthy	15 153 300	67.49	94.06	71.75	255.7
F01-2	Sick	15 292 600	67.91	93.49	72.63	259.6
F02-1	Healthy	15 322 500	66.75	93.62	71.29	255.7
F02-2	Sick	15 865 300	65.79	93.97	70.01	260.9
F03-1	Healthy	20 305 100	67.78	93.73	72.31	344
F03-2	Healthy	19 558 200	66.35			324.4
F03-3	Sick	21 670 000	69.08			374.2
F04-1	Healthy	15 425 900	67.84	93.82	72.30	261.6
F04-2	Healthy	15 555 000	70.34	94.27	74.61	273.5
F04-3	Sick	15 206 100	71.07	94.32	75.34	270.1

Enrichment of the targeted region by 700 fold

Sequencing QC2

Do we sequence the entire region ?

Samples	Statut	3Mb Target Region Coverage (%)	3Mb Region Average Depth (X)	1.6 Mb Target Region Coverage (%)	1.6 Mb Region Average Depth (X)
F01-1	Sain	81.40	301.48	100	301.38
F01-2	Atteint	82.83	300.86	100	300.89
F02-1	Sain	82.64	296.98	100	297.09
F02-2	Atteint	82.40	303.97	99.98	303.91
F03-1	Sain	83.33	289.08	100	289.28
F03-2	Sain	83.85	282.44	100	282.46
F03-3	Atteint	88.00	277.02	100	278.16
F04-1	Sain	82.77	303.37	100	303.58
F04-2	Sain	82.33	318.88	100	319.17
F04-3	Atteint	82.05	316.07	100	316.01

Sequencing QC3
Is the coverage homogeneous among the region?

Individus	P10/Région (3Mb)	P10/Région (1.6Mb)	P5/Region (3Mb)	P5/Région (1.6Mb)	P1/Region (3Mb)	P1/Région (1.6Mb)
F01-1	71.57	83.80	75.56	88.46	82.07	96.06
F01-2	71.47	82.86	75.65	87.08	82.54	94.99
F02-1	71.45	82.40	76.14	87.79	83.85	96.68
F02-2	71.33	82.43	75.74	87.53	82.33	95.14
F03-1	74.34	85.08	77.62	88.84	84.27	96.43
F03-2	74.39	84.57	77.91	88.58	84.71	96.31
F03-3	73.27	79.32	77.70	84.12	87.70	94.96
F04-1	71.71	82.65	75.98	87.53	82.53	95.09
F04-2	71.12	82.39	75.50	87.45	82.01	94.99
F04-3	71.20	82.72	75.52	87.73	82.08	95.35

- P10: 30x, P5: 15x, P1: 3x
- We probably need 15x

		F01-2								
		Homo	Homo.Douteux	Homo.SNPcomplex	Homo.ref	HTZ	HTZ.Douteux	HTZ.SNPcomplex		
F01-1	Homo	599	0	1	343	0	0	0	943	1313
	Homo.Douteux	0	0	0	2	1	0	0	3	
	Homo.SNPcomplex	4	0	2	1	0	0	0	7	
	Homo.ref	354	3	0	0	3		0	360	
	HTZ	310	0	2	1014	88	6	5	1425	1519
	HTZ.Douteux	4	0	0	40	1	0	0	45	
	HTZ.SNPcomplex	3	0	3	20	2	0	20	48	
	HTZ.SNPcomplex.Douteux	0	0	0	1	0	0	0	1	
		1274	3	8	1421	95	6	25		
		2706				126				

Exome Sequencing

SureSelect™ Target Enrichment System:

Human All Exons in a Tube

- 38 Mb: CCDS + >1,000 ncRNA
- 1 sample = 1 tube = 1 lane (2 x 76bp)
- No gel library preparation!
- 3ug starting gDNA

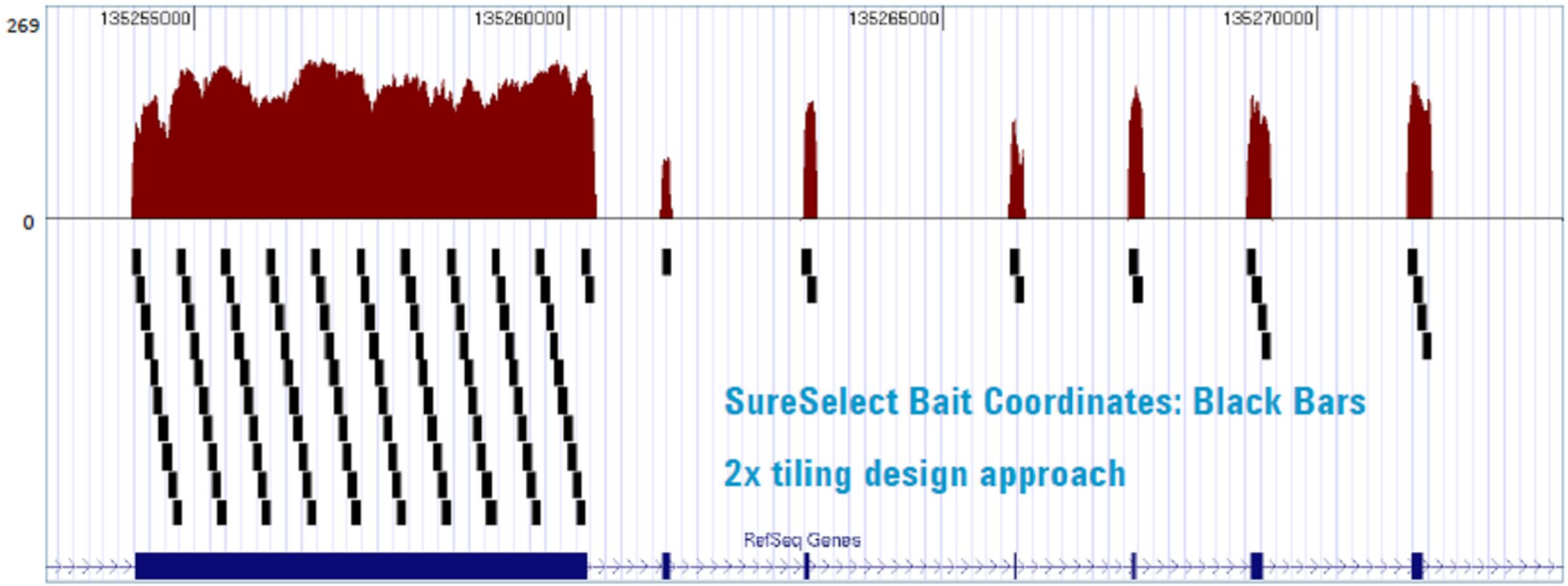
Available October 1, 2009

Parameters

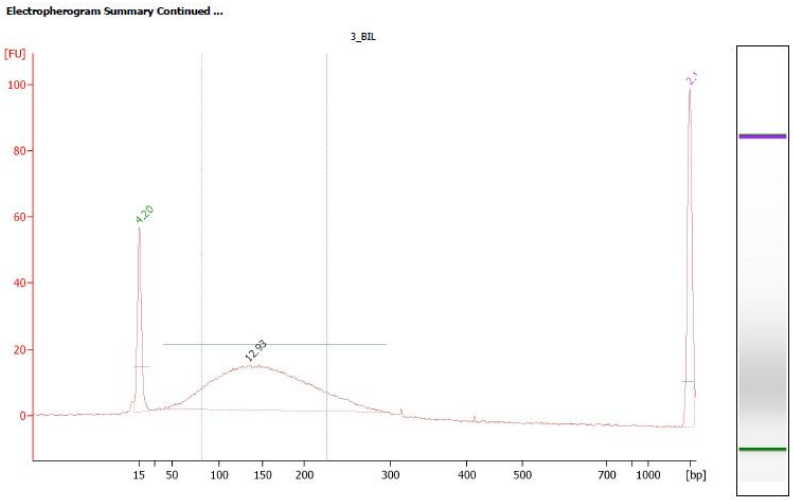
- HapMap Samples used for development
- Target : 38 Mb
- Number of exons : ~180 000 (CCDS Database) + 1000 nc exons
- 120-mer baits, end to end tiled
- 3 µg starting material
- 2x75 bp PE sequencing

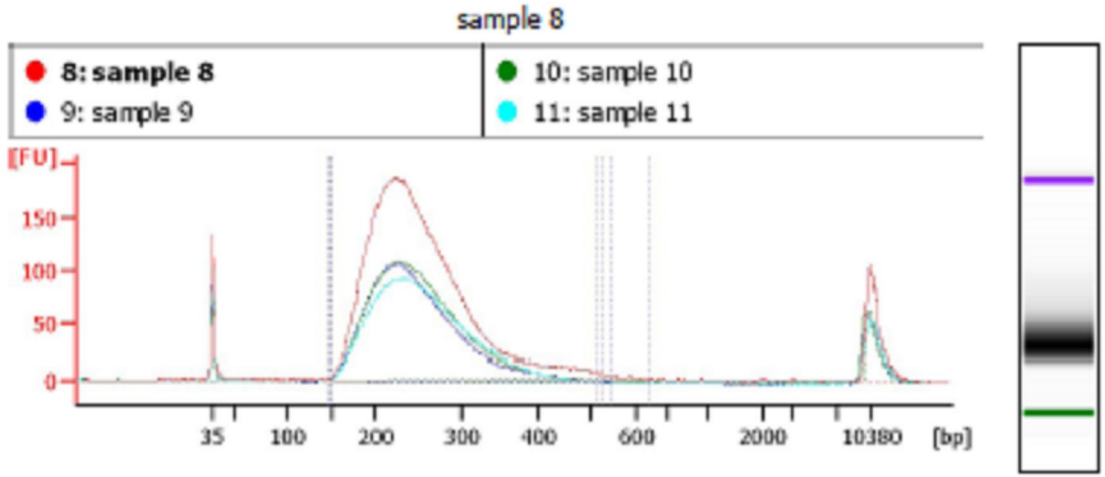
- Min guaranteed :
  - 18 M cluster, 2,7 Gb
  - Expected 70% in Target
  - Avg coverage : 50X

Capture Design : 2X tiling

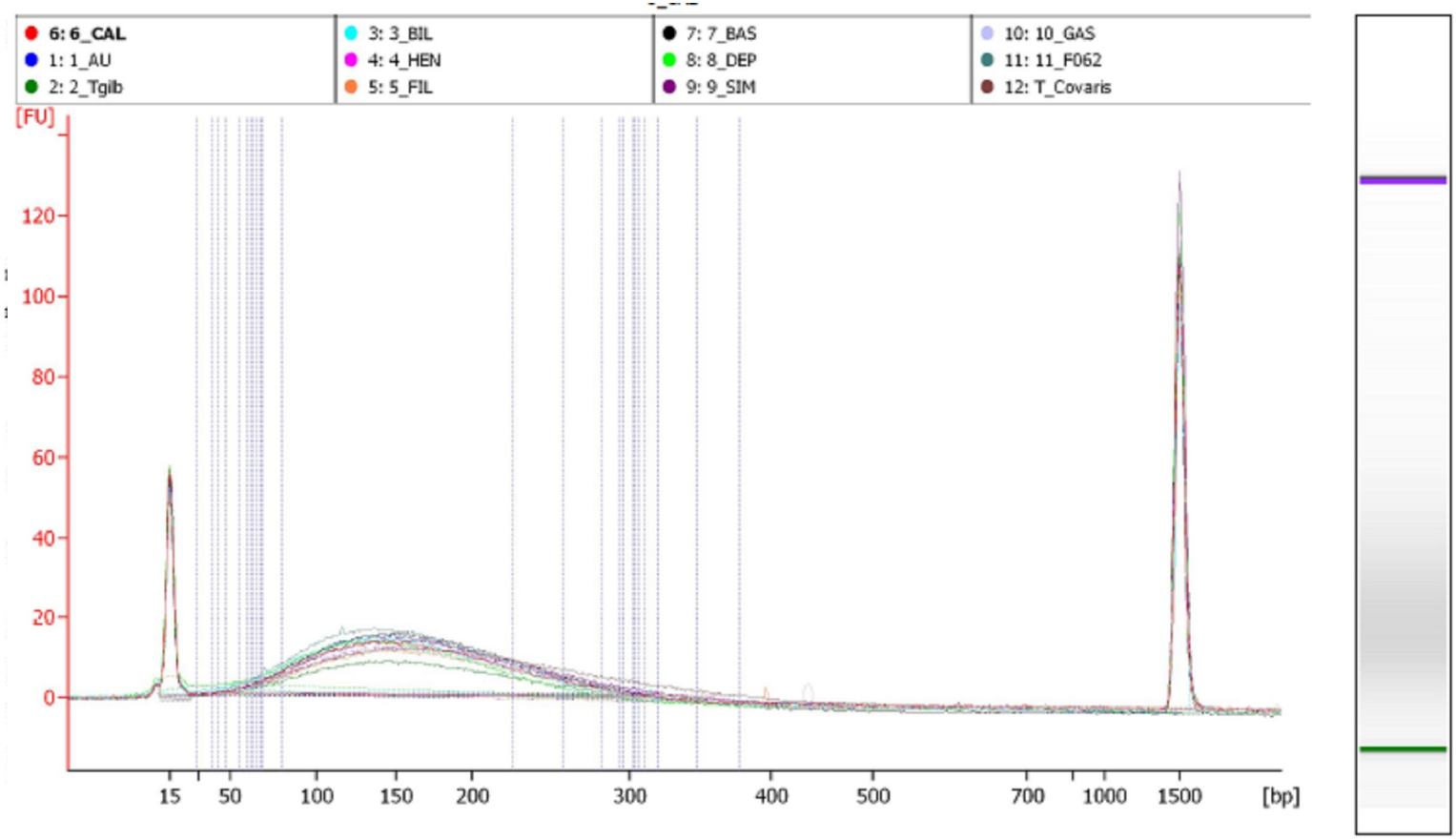


Fragmentation focused at 150 bases





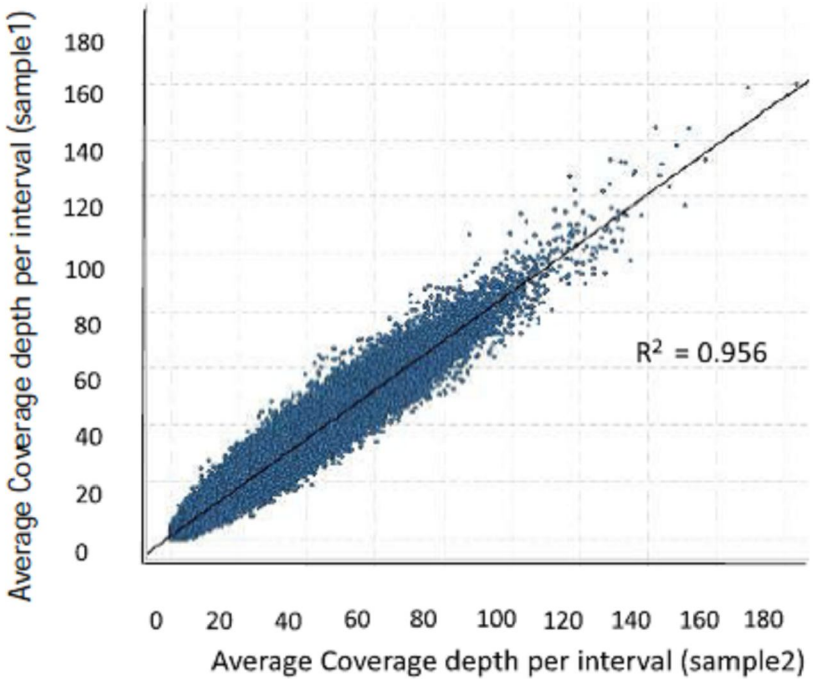
Fragmentation for 11 libraries



Metrics

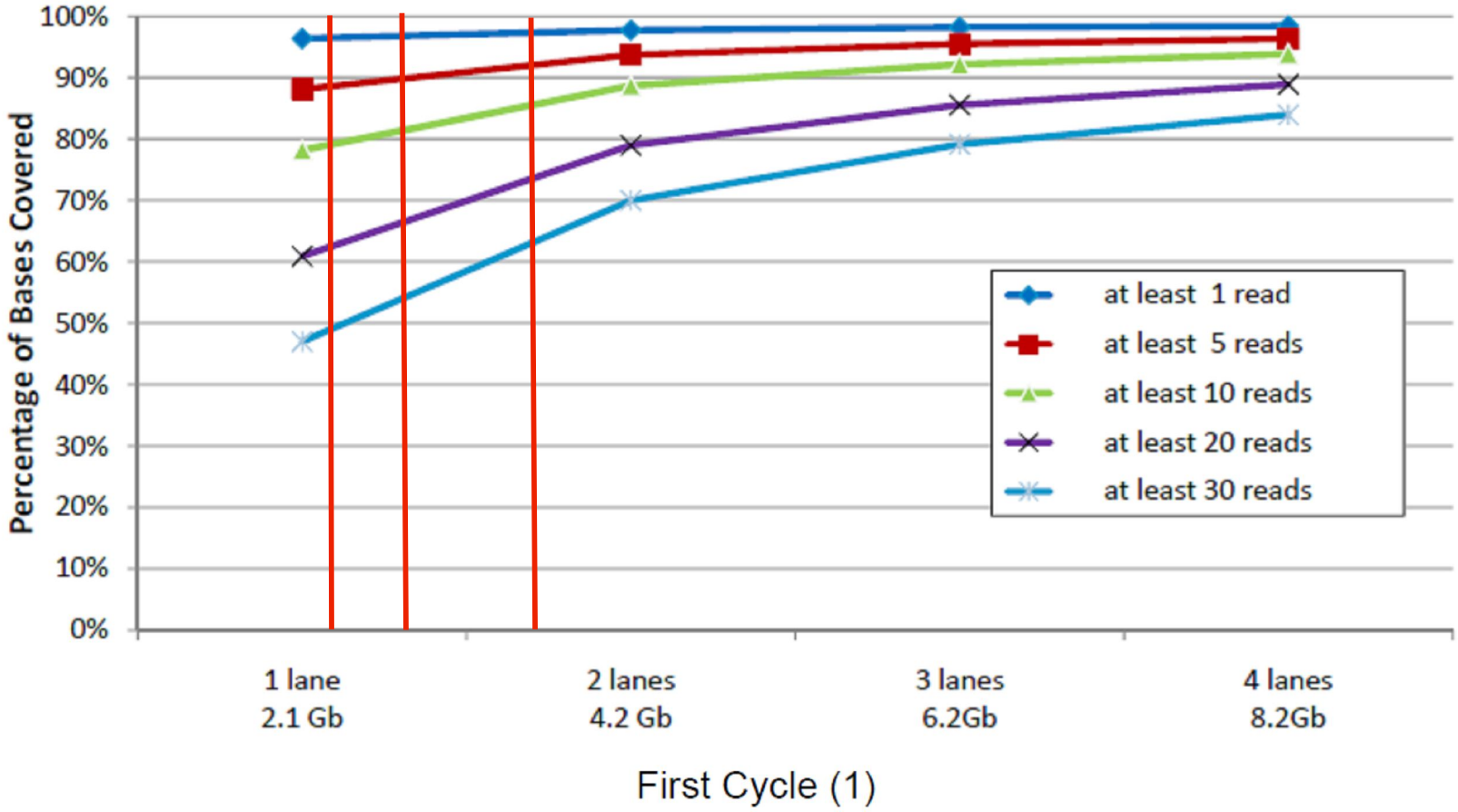
- Specificity: % of reads that map to targeted regions
- Coverage: % of bases with 1, 5, 10, 20, 40x coverage
- Comparison vs. HapMap sensitivity: % of known SNP called
- Comparison vs. dbSNP

Human All Exon in a tube: High reproducibility



- Target: 38 Mb
- Exons targeted: ~180,000 (CCDS database)
- +700 miRNA (Sanger v13)
- + 300 ncRNA
- 120-mer baits, 1x tiling
- 1 tube/capture/lane of Illumina Sequencer (2x75bp)

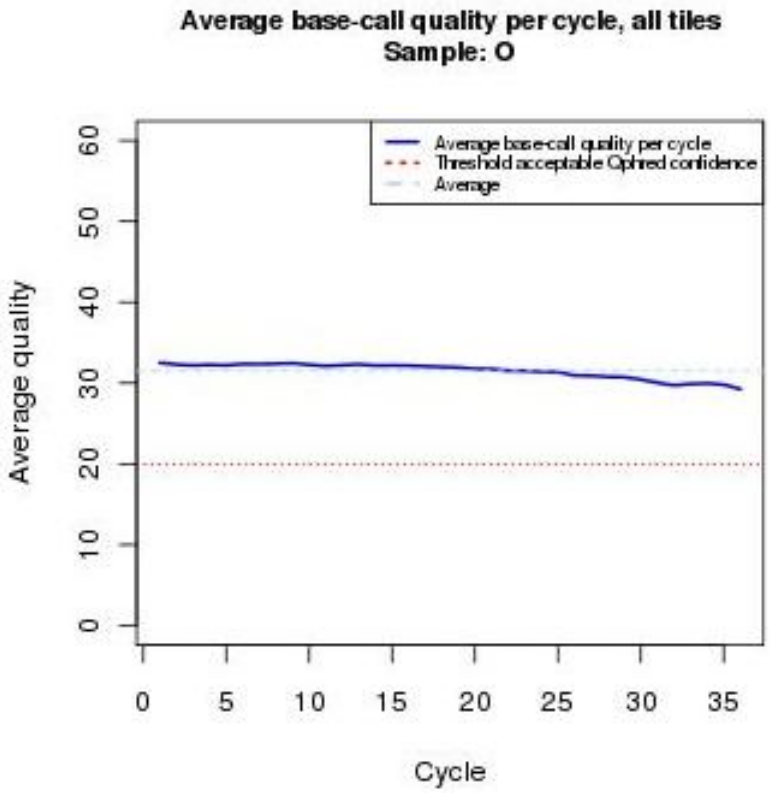
What will be the missing data?

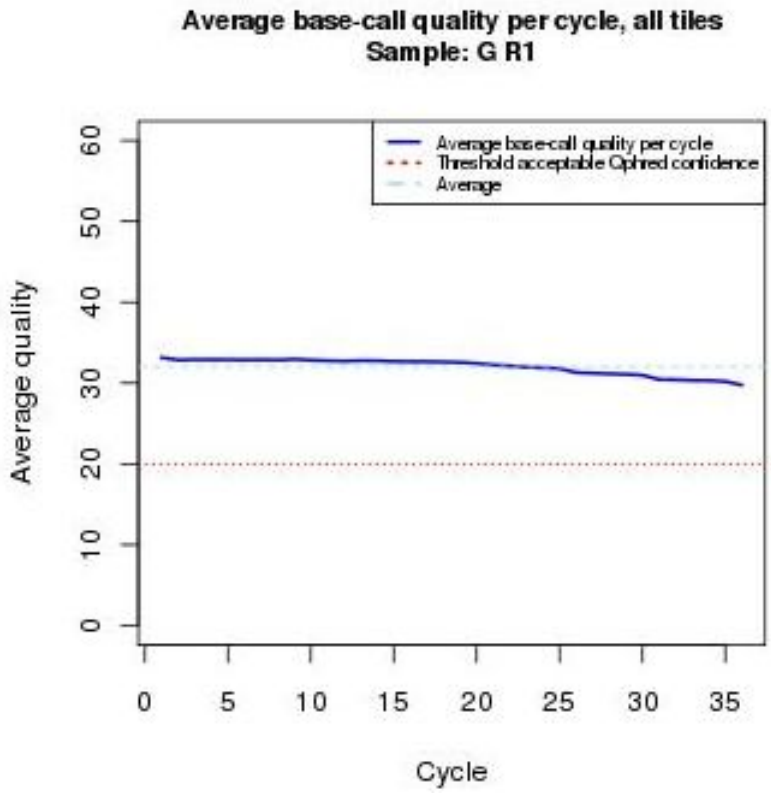


MetricName	Lane1	Lane2	Lane3	Lane4	Lane5	Lane6	Lane7	Lane8
# of Clusters	285,133	277,583	276,973	277,968	288,878	277,560	282,262	266,666

Bioinformatics

Quality Control



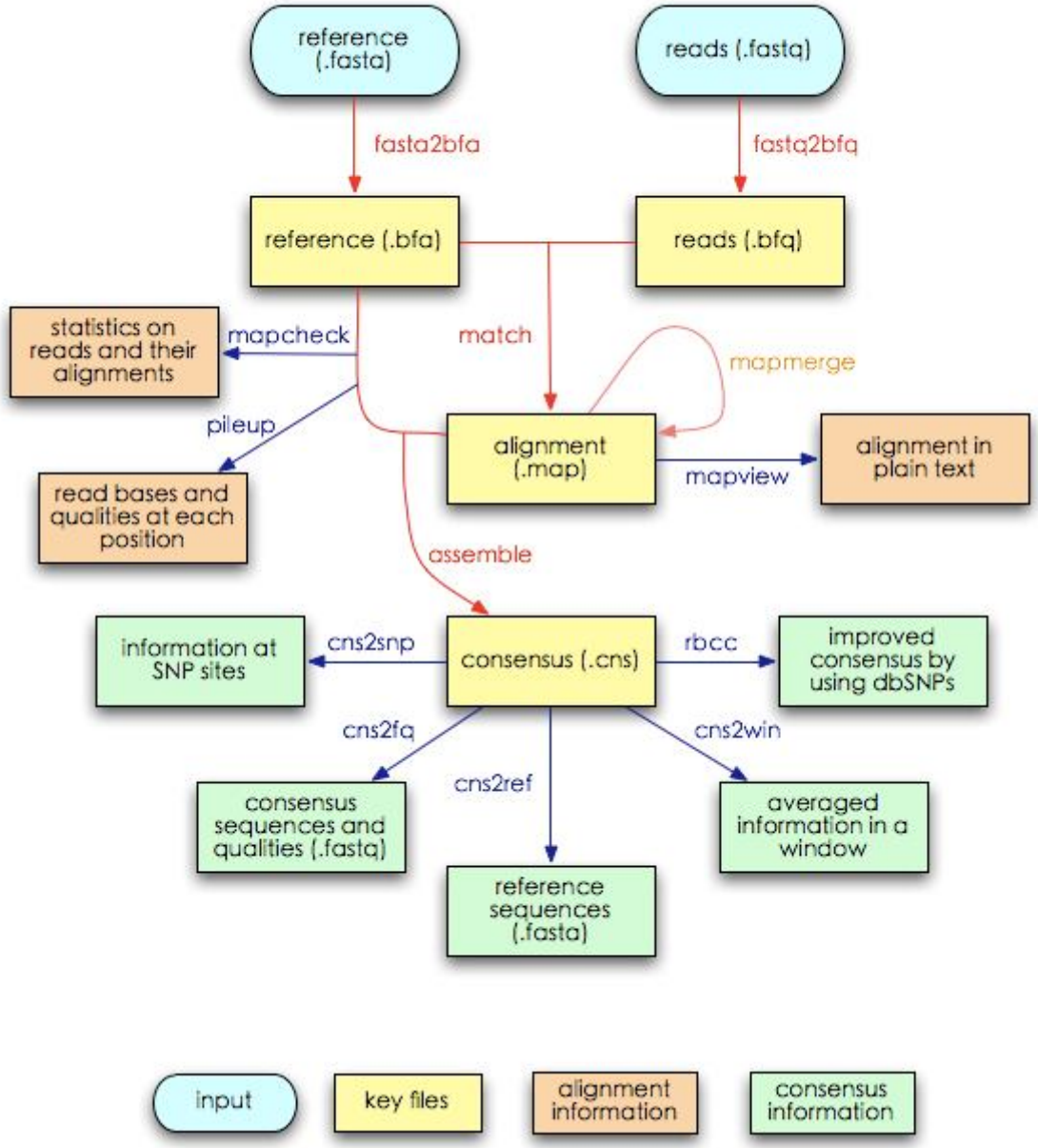


- Qphred by cycle, all tiles

Bioinformatic Analysis

- Illumina Pipeline GA IIx (CASAVA1.6)
  - Image Analysis
  - Base calling
  - Retrieving fastq files
- Alignment using MAQ
  - align against ref sequence → generate a consensus → detection des SNPs et Indels
  - Parameters
    - -n : max mismatches per read
    - -r: heterozygote fraction
    - -a: max insert size
    - -q: base quality
    - Eliminate identical pairs (PCR bias)

MAq Overview



Our pipeline for SNPs detection

Out.pileup

Ref/chrom	pos	base ref	depth	read base
hg18_dna	1	C	0	@
hg18_dna	2	T	2	@ ,.
hg18_dna	3	T	14	@ ,.,,,,,,,,,,,.

….

hg18_dna 5537 C 207

@,,TTTTttt.....tttT,,T,.tttt..tttt.ttT..t..t.tt.tttttt,,,tt..t,T,,,T,TtT,tT.t....tT.tttt.t,.t...tt.t..,TT, T.ttTT,T,,T,,.t,,T.TtttT..ttttt...t.ttt,TTTTTTTT,,.TT,Ttt.,.ttt..tt,T,Ttt.t.tttT,.ttt,.....,,T,ttt...

Reads bases identical to reference = , (forward) or . (Reverse)
Reads bases differents from the reference = letters upper (forward) / lower (Reverse)

Allelic Frequence SNPs = nbre(, et .) / depht
SNP Depth = depht

- → N**ucleotide Count**
- → **Define a consensus using statistics**
- → **Give a confidence value (Score) to genotypes**
- → **Score Cut-Off**

Genotype Call

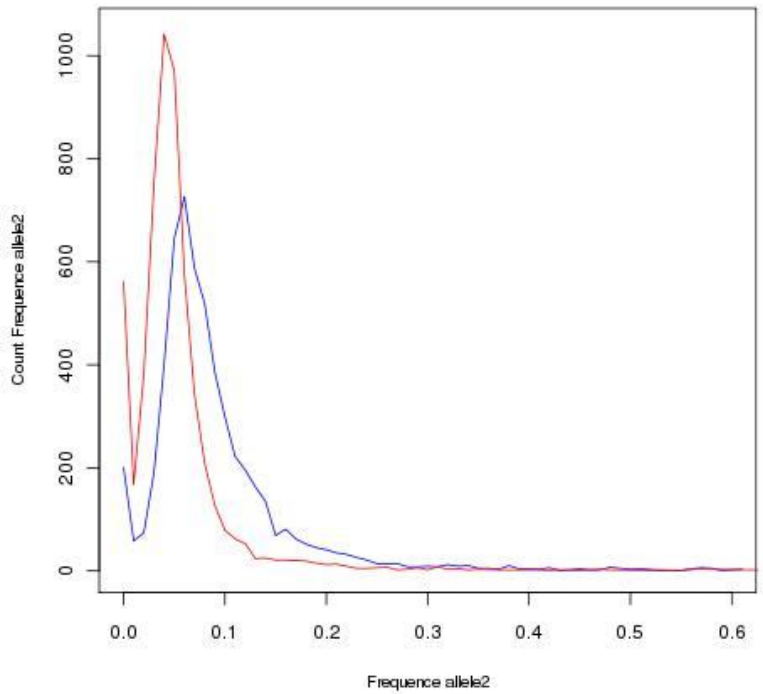
$$P(<AA>|\ D) = \binom{n}{k} * \xi^{n-k} * (1-\xi)^{k} * \frac{(1-r)}{2}$$

$$P(<BB>|\ D) = \binom{n}{k} * \xi^{k} * (1-\xi)^{n-k} * \frac{(1-r)}{2}$$

$$P(<AB>|\ D) = \frac{\binom{n}{k}}{2^{n}} * r$$

n : Depth
k : allele count A or B
ε : error rate
r : heterozygote fraction

Error Rate



Results

- 5 categories (Cut Off : 1000)
  - Homo: Homozygous SNP → a non-reference allele is observed
  - Homo.Douteux: Homozygous SNP with low confidence
  - HTZ: Heterozygous SNP
  - HTZ.Douteux: Heterozygous SNP with low confidence
  - Homo.ref: non SNP, Homozygous → a reference allele is observed
- What do we give?
  - F01_SNPDetectionTable.xls
  - ..\SNPsDetectedOurAlgoMAQ.xls

Additionnal Annotation

- In order to give you really analyzed data

Nom SNP	Chrom	Position	Gene name	Type	N°exon	codon wild	codon mut	aaw ild	aa mut	wild prot	mut prot
rs1	1	1116	uc001aaa.2	3'UTR							
rs1	1	1116	uc009vip.1	3'UTR							
rs2	1	1149167	uc009vjv.1	exon	2	CAG	TAG	Q	*	MQRWIMEKTAEHFQE	M*RWIMEKTAEHFQE
rs3	1	1205906	uc001adt.1	exon	1	ATG	ATA	M	I	MRAVLSQKTTPLPRYL	IRAVLSQKTTPLPRYL
rs3	1	1205906	uc001adu.1	5'UTR							

Who acts in IntegraGen?

Dr Bernard Courtieu
CEO

Patricia Lewin
VP R&D & Medical Director

Patrick Court
CFO & COO

Larry Yost
VP- US Operations

Emmanuel Martin
CCO, Head of Services & Oncology

Francis Rousseau
Director of Genomics

And the entire Lab Team