IntegraGen at a glance

The n°1 privately-owned genomics platform in France

Autism

Oncology

Genomics Services
Serves the researcher’s most complex needs in genomics

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

Genomic DNA library \rightarrow \text{Hybridization} \rightarrow \text{Elution} \rightarrow \text{Sequence} \rightarrow \text{Enrichment Oligos library}
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: 120 bases
  - Baits coverage: 5X
  - Baits overlap: 24 bases
  - Centered Design
  - % Region Covered: 50.81%
  - Effective region size: 1,609 Mb

- **Masked regions:** 49.19%

- **Sequence:**
  - Fragment size: 400 bases
  - 1 sample by Flow-Cell lane
  - Single Read 75 bases

![Depth distribution on sequence](chart.png)
Sequence coverage represented by start positions
Sequencing QC1
How does the enrichment work?

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

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

Enrichment of the targeted region by 700 fold
Sequencing QC2

Do we sequence the entire region?

<table>
<thead>
<tr>
<th>Samples</th>
<th>Statut</th>
<th>3Mb Target Region Coverage (%)</th>
<th>3Mb Region Average Depth (X)</th>
<th>1.6 Mb Target Region Coverage (%)</th>
<th>1.6 Mb Region Average Depth (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>F01-1</td>
<td>Sain</td>
<td>81.40</td>
<td>301.48</td>
<td>100</td>
<td>301.38</td>
</tr>
<tr>
<td>F01-2</td>
<td>Atteint</td>
<td>82.83</td>
<td>300.86</td>
<td>100</td>
<td>300.89</td>
</tr>
<tr>
<td>F02-1</td>
<td>Sain</td>
<td>82.64</td>
<td>296.98</td>
<td>100</td>
<td>297.09</td>
</tr>
<tr>
<td>F02-2</td>
<td>Atteint</td>
<td>82.40</td>
<td>303.97</td>
<td>99.98</td>
<td>303.91</td>
</tr>
<tr>
<td>F03-1</td>
<td>Sain</td>
<td>83.33</td>
<td>289.08</td>
<td>100</td>
<td>289.28</td>
</tr>
<tr>
<td>F03-2</td>
<td>Sain</td>
<td>83.85</td>
<td>282.44</td>
<td>100</td>
<td>282.46</td>
</tr>
<tr>
<td>F03-3</td>
<td>Atteint</td>
<td>88.00</td>
<td>277.02</td>
<td>100</td>
<td>278.16</td>
</tr>
<tr>
<td>F04-1</td>
<td>Sain</td>
<td>82.77</td>
<td>303.37</td>
<td>100</td>
<td>303.58</td>
</tr>
<tr>
<td>F04-2</td>
<td>Sain</td>
<td>82.33</td>
<td>318.88</td>
<td>100</td>
<td>319.17</td>
</tr>
<tr>
<td>F04-3</td>
<td>Atteint</td>
<td>82.05</td>
<td>316.07</td>
<td>100</td>
<td>316.01</td>
</tr>
</tbody>
</table>
### Sequencing QC3

Is the coverage homogeneous among the region?

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

- **P10: 30x, P5: 15x, P1: 3x**

- **We probably need 15x**
<table>
<thead>
<tr>
<th></th>
<th>Homo</th>
<th>Homo.Douteux</th>
<th>Homo.SNPcomplex</th>
<th>Homo.ref</th>
<th>HTZ</th>
<th>HTZ.Douteux</th>
<th>HTZ.SNPcomplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo</td>
<td>599</td>
<td>0</td>
<td>1</td>
<td>343</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Homo.Douteux</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Homo.SNPcomplex</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Homo.ref</td>
<td>354</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTZ</td>
<td>310</td>
<td>0</td>
<td>2</td>
<td>1014</td>
<td>88</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>HTZ.Douteux</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>40</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTZ.SNPcomplex</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>20</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HTZ.SNPcomplex .Douteux</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

|          | 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

SureSelect Bait Coordinates: Black Bars

2x tiling design approach
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?
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)
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......ttttT,,T,,ttt..tttt.ttT..t...ttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
Genotype Call

\[ P(<AA>|D) = \binom{n}{k} \cdot \xi^{n-k} \cdot (1-\xi)^k \cdot \frac{(1-r)}{2} \]
\[ P(<BB>|D) = \binom{n}{k} \cdot \xi^k \cdot (1-\xi)^{n-k} \cdot \frac{(1-r)}{2} \]
\[ P(<AB>|D) = \frac{\binom{n}{k}}{2n} \cdot r \]

\( n \) : Depth
\( k \) : allele count A or B
\( \varepsilon \) : 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

<table>
<thead>
<tr>
<th>Nom SNP</th>
<th>Chrom</th>
<th>Position</th>
<th>Gene name</th>
<th>Type</th>
<th>N°exon</th>
<th>codon wild</th>
<th>codon mut</th>
<th>aa w/ld</th>
<th>aa mut</th>
<th>wild prot</th>
<th>mut prot</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1</td>
<td>1</td>
<td>1116</td>
<td>uc001aa.2</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1</td>
<td>1</td>
<td>1116</td>
<td>uc009vlp.1</td>
<td>3'UTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2</td>
<td>1</td>
<td>1149167</td>
<td>uc009vju.1</td>
<td>exon 2</td>
<td></td>
<td>CAG</td>
<td>TAG</td>
<td>Q</td>
<td>*</td>
<td>MQRWIMEKTAEHFQE</td>
<td>M*RWIMEKTAEHFQE</td>
</tr>
<tr>
<td>rs3</td>
<td>1</td>
<td>1205906</td>
<td>uc001adt.1</td>
<td>exon 1</td>
<td></td>
<td>ATG</td>
<td>ATA</td>
<td>M</td>
<td>I</td>
<td>MRAVLSQKTTPLPRYL</td>
<td>IRAVLSQKTTPLPRYL</td>
</tr>
<tr>
<td>rs3</td>
<td>1</td>
<td>1205906</td>
<td>uc001adu.1</td>
<td>5'UTR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Who acts in IntegraGen?

Dr Bernard Courtieu  
CEO

Patrick Court  
CFO & COO

Emmanuel Martin  
CCO, Head of Services & Oncology

Patricia Lewin  
VP R&D & Medical Director

Larry Yost  
VP - US Operations

Francis Rousseau  
Director of Genomics

And the entire Lab Team