A transcriptional sketch of a human breast cancer by 454 deep sequencing
The cancer transcriptome is difficult to explore, due to the heterogeneity of quantitative and qualitative transcriptional events linked to the disease status.

An increasing number of “unconventional” transcripts, such as novel isoforms, noncoding RNAs, somatic gene fusions and deletions have been associated with the tumoral state.

Massively parallel sequencing techniques make full-transcriptome sequencing feasible with a limited laboratory and financial effort, and provide a framework for exploring the complexity inherent to the cancer transcriptome.
We developed a 454 deep sequencing and bioinformatics analysis protocol to investigate the molecular composition of a breast cancer poly(A)+ transcriptome.

This method utilizes a normalization step to diminish the abundance of highly expressed transcripts and biology-oriented bioinformatic analyses to facilitate detection of rare and novel transcripts, which may enhance our understanding of the aetiology of the disease.

We demonstrate that combining 454 deep sequencing with a normalization step and careful bioinformatic analysis facilitates the discovery of rare transcripts, and can be used as a qualitative tool to characterize transcriptome complexity, revealing many hitherto unknown transcripts, splice isoforms, and gene fusion events, even at a relatively low sequence sampling.
# Library normalization: wet lab matters

## References

<table>
<thead>
<tr>
<th>Reference Gene</th>
<th>454 Reads mapped to the genome (194,806)</th>
<th>UniGene ESTs (39,700)</th>
<th>Probability of differential expression between the libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTB</td>
<td>11</td>
<td>187</td>
<td>Prob &gt; 0.999</td>
</tr>
<tr>
<td>GAPDH</td>
<td>31</td>
<td>225</td>
<td>Prob &gt; 0.999</td>
</tr>
<tr>
<td>HPRT1</td>
<td>7</td>
<td>0</td>
<td>0.5 &lt; Prob &lt; 0.6</td>
</tr>
</tbody>
</table>
(left) Distribution of sequence lengths shows a good approximation to a Normal Distribution

(right) Sequence reads show an higher representation toward the 3’end of a transcript, but coverage is present along all the transcript
HPC challenges (and solutions)

(1) **Data distribution**: split the dataset on chunks of, e.g., 2,000 sequences each.

(2) **Memory management** may be very problematic with short word search parameters (-W 4 for Blast or – tileSize 8 for Blat) when comparing against the large and repetitive human or mouse genomes.

(3) String-based searches eat up lots of disk space and processor time quickly

=> we used two bioinformatic clusters and an eight-processor server with large shared memory (8 Giga)
### Mapping to the Genome and the Transcriptome

<table>
<thead>
<tr>
<th>Set Description</th>
<th>Number of reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (unfiltered)</td>
<td>251,262</td>
</tr>
<tr>
<td>Mapping to the genome, 70% coverage, high stringency</td>
<td>194,806</td>
</tr>
<tr>
<td><strong>Subset with a single match on the genome at 98% identity and 98% coverage (98.98.1 dataset)</strong></td>
<td><strong>132,113 reference dataset</strong></td>
</tr>
<tr>
<td>Subset with a single match on the genome and 100% coverage of the alignment</td>
<td>114,427 – 87% of the reference dataset</td>
</tr>
<tr>
<td>Subset of 98.98.1 dataset matching with max 6 errors (mismatches + indels) and 90% coverage on UCSC all_mrna and RefSeq – canonical transcripts dataset</td>
<td>59,632 - – 45% of the reference dataset</td>
</tr>
<tr>
<td><strong>Subset of 98.98.1 dataset matching inside an UCSC Known Gene (Intragenic dataset, intronic + exonic transcripts)</strong></td>
<td><strong>118,840 - – 90 % of the reference dataset</strong></td>
</tr>
<tr>
<td>Matching with max 6 errors (mismatches + indels) and 90% coverage to the Human ORESTES EST dataset (764,587 sequences)</td>
<td>68,396 – 52% of the reference dataset</td>
</tr>
</tbody>
</table>
## Classification of the sequences in the genome annotation context

<table>
<thead>
<tr>
<th>Sequence class</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of reads</strong></td>
</tr>
<tr>
<td>Intergenic Unspliced</td>
</tr>
<tr>
<td>6.298</td>
</tr>
<tr>
<td>Intergenic Spliced</td>
</tr>
<tr>
<td>402</td>
</tr>
<tr>
<td>Intragenic Unspliced – total</td>
</tr>
<tr>
<td>97.690</td>
</tr>
<tr>
<td>3 TERM</td>
</tr>
<tr>
<td>2.475</td>
</tr>
<tr>
<td>(Poli-A)</td>
</tr>
<tr>
<td>(989)</td>
</tr>
<tr>
<td>(INTERNAL)</td>
</tr>
<tr>
<td>(1.486)</td>
</tr>
<tr>
<td>5 TERM</td>
</tr>
<tr>
<td>2.807</td>
</tr>
<tr>
<td>(TSS)</td>
</tr>
<tr>
<td>(1.113)</td>
</tr>
<tr>
<td>(INTERNAL)</td>
</tr>
<tr>
<td>(1.694)</td>
</tr>
<tr>
<td>EXON</td>
</tr>
<tr>
<td>1.331</td>
</tr>
<tr>
<td>INTRAEXON</td>
</tr>
<tr>
<td>64.326</td>
</tr>
<tr>
<td>INTRON</td>
</tr>
<tr>
<td>26.751</td>
</tr>
<tr>
<td>Intragenic Spliced</td>
</tr>
<tr>
<td>10.037</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>114.427</td>
</tr>
</tbody>
</table>

**Abbreviations:**
- 3 TERM, read which extend the annotated 3’ term of the target gene.
- Poli-A: read which extends at 3’ the last exon.
- INTERNAL: read which extends at 3’ any exon except the last.
- 5 TERM: read which extend the annotated 5’ term of the target gene.
- TSS: read which extends at 5’ the first exon.
- INTERNAL: read which extends at 5’ any exon except the first.
- EXON: read mapping inside an exon with one or both ends coincident with exon boundaries.
- INTRAEXON: read mapping completely inside an exon of the target gene.
- INTRON: read mapping completely inside an intron.
Genes or Transcripts Fusions in solid tumors

Nature Reviews Cancer 7, 233-245 (April 2007)

The impact of translocations and gene fusions on cancer causation

Felix Mitelman- Bertil Johansson-& Fredrik Merten

Abstract

Chromosome aberrations, in particular translocations and their corresponding gene fusions, have an important role in the initial steps of tumorigenesis; at present, 358 gene fusions involving 337 different genes have been identified. An increasing number of gene fusions are being recognized as important diagnostic and prognostic parameters in malignant haematological disorders and childhood sarcomas. The biological and clinical impact of gene fusions in the more common solid tumour types has been less appreciated. However, an analysis of available data shows that gene fusions occur in all malignancies, and that they account for 20% of human cancer morbidity. With the advent of new and powerful investigative tools that enable the detection of cytogenetically cryptic rearrangements, this proportion is likely to increase substantially.
Genes or Transcripts Fusions

Perfect Fusion of two sequences located in different chromosomes
**Genes or Transcripts Fusions**

**UBR4**, commonly known as p600 or retinoblastoma protein-associated factor 600, is a cellular target of the human papillomavirus type 16 E7 oncoprotein that regulates cellular pathways, contributing to anchorage-independent growth and cellular transformation. UBR4-E7 interaction strongly contributes to cellular transformation (Huh 2005).

The **GLB1** gene encodes beta-galactosidase-1 (EC 3.2.1.23), a lysosomal hydrolase that cleaves the terminal beta-galactose from ganglioside substrates and other glycoconjugates. The predicted fusion, verified by direct sequencing of the original cDNA library, links exon 16 of the gene **UBR4** with the terminal exon (coding + 3’UTR), common to all the transcript variants, of the **GLB1** gene.

Our sequence (4A) is colinear with both transcripts and exon-exon junctions are clear in the hybrid sequence. The predicted final processed fusion cDNA **UBR4/GLB1** would be 14,022-bp long and would produce a very large protein of 4,526 residues, which however is shorter than the original UBR4 protein (5,183 residues).
A: Chr1:19298680-19298756. 16th exon from the 3' end of the UBR4 gene.

00000001 tctagtgaacaataaatacattagttttgtaccttcctgtggctgaagtt 00000050
<<<<<<<<<<<|……………………………………………………………………………………………………………………|<<<<<<<<<<<
19298756 tctagtgaacaataaatacattagttttg.accttcctgtggctgaagtt 19298708

00000051 tacaagaaagtctggtgtaccacgaatg 00000078
<<<<<<<<<<<|……………………………………………………………………………………………………………………|<<<<<<<<<<<
19298707 tacaagaaagtctggtgtaccacgaatg 19298680

B. Chr3:33013785-33013825. Exon 15 (last) of the GLB1 gene.

00000075 aatggttttaaccttggccgctattggccagccgcccccccc 00000115
<<<<<<<<<<<|……………………………………………………………………………………………………………………|<<<<<<<<<<<
33013825 aatggttttaaccttggccgctattggccagccgcccccccc 33013785

C.

<table>
<thead>
<tr>
<th>4401</th>
<th>4450</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBR4-GLB1 (4401)</td>
<td>EDDSGMELLVNNKIIISLDLPVAEVYKKVWCTTN</td>
</tr>
<tr>
<td>GLB1 (4401)</td>
<td>EDDSGMELLVNNKIIISLDLPVAEVYKKVWCTTNE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4451</th>
<th>4500</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBR4-GLB1 (4451)</td>
<td>FVPQHILMTSAAPTITVLELEWACPSSDDPELCAVTFDREMGSSVTYD</td>
</tr>
<tr>
<td>GLB1 (4451)</td>
<td>ATEEFIESLDSTTDEEDEEEVYKMAVQCGGILECMNLRTAGIRDFKQ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4501</th>
<th>4541</th>
</tr>
</thead>
<tbody>
<tr>
<td>UBR4-GLB1 (4501)</td>
<td>HPSKPVEKRLMPPPPQKNDLSWLDHV</td>
</tr>
<tr>
<td>GLB1 (4501)</td>
<td>GRHLLTVLKLFSYCVKVKVNRQQLVKLEMNLTNVMLGTLN</td>
</tr>
</tbody>
</table>
An example of a transcriptional or genomic deletion event is provided by the sequence read 1B (167378_1645_3303), located on chromosome 8. We interpret this sequence as a deletion, probably due to a loop which causes the inclusion of exons 2 and 7 of the \textit{WHSC1L1} gene in an inverted order in the mature transcript. This transcript was also confirmed by direct sequencing of the cDNA library.

The \textit{WHSC1L1} gene is related to the Wolf-Hirschhorn syndrome candidate-1 gene and encodes a protein with PWWP (proline-tryptophan-tryptophan-proline) domains. Two alternatively spliced \textit{WHSC1L1} variants have been described. The long isoform contains a PHD-finger domain (an interleaved type of Zn finger chelating 2 Zn ions) and a SET domain (protein-protein interaction domain); however, the function of the protein has not been determined yet and hence the relevance to cancer aetiology of this deletion is uncertain.
A. chr8:38303404-38303449. WHSC1L1 exon # 7

```
00000010 caacgcagagtttgtaatctctctgaagcaacatctggtctacag
<<<<<<<<<<< 00000057
<<<<<<<<<<<
38303449 caacgcagagatgtatcatctctgaacatctggtctacag
38303404
```

B. chr8:38324848-38324892. WHSC1L1 exon # 2

```
00000055 aggcocgtgatgaggaagggaaagttaagggtgctggagcaaac
<<<<<<<<<<< 00000099
<<<<<<<<<<<
38324892 aggcocgtgatgaggaagggaaagttaagggtgctggagcaaac
38324848
```

C.
Rare and Novel Isoforms

The following examples clearly demonstrate that it is possible to identify known, novel and even possible cancer-related isoforms with this sequence length and at this sequencing depth.

We were able to retrieve known isoforms of *IGL* (Unigene cluster 449585 Immunoglobulin lambda joining 3), which is located in an area of very active and complex genomic rearrangements. The corresponding read 045624_1590_1179 (which we renamed 6B) is a 102 nt transcript fragment which maps to the Variable, Light and Join segments of immunoglobulin genes on Chr 22q11.1-q11.2.

When mapped to the genome this read aligns to the first 60 nucleotides, with 98% identity, to nt 21571798-21573177 (divided in two exons) and with nt 60-101, with 95% identity, to nt 21060830-21060871 (third exon) of the minus strand of Chr 22.

Hence this single 102 nt read contains three different exons, the second and the third separated by around 511,000 bases.
A. Chr11:82460863-82461422

00000001 tcttcaggctattccctctctcttttaattgacaggccgaagcaacgt 00000050
<<<<<<<<<
82461422 tcttcaggctattccctctctcttttaattgacaggccgaagcaacgt 82461373

00000051 cttttacacttggtcactttcacaacagattttctccgacgctttg 00000100
<<<<<<<<<
82461372 cttttacacttggtcactttcacaacagattttctccgacgctttg 82461323

00000101 ttaatcattc 00000110
<<<<<<<<<
82461322 ttaatcattc 82461313

00000111 cgccccac 00000119
<<<<<<<<<
82460871 cgccccac 82460863

B.
Cancer-associated isoforms have been usually predicted on the basis of differential representation of ESTs between tumor and healthy tissues. A full genome survey of our 454 cDNA sequence reads against the latest version of the ASAPII database identified five putative cancer-associated isoforms, one of which has a significant biological plausibility: the isoform b or shorter isoform of the \textit{HIGD1A} gene (HIG1 domain family, member 1A) gene which encodes a protein containing a domain found in proteins thought to be involved in the response to hypoxia.

Growth and progression of breast cancers are accompanied by increased neovascularization (angiogenesis). A variety of factors, including hypoxia and genetic changes in tumor cells, contribute to increased production of angiogenic factors (Boudreau 2003).
<table>
<thead>
<tr>
<th>Internal ID</th>
<th>Read ID</th>
<th>Primer For (5' =&gt; 3')</th>
<th>Primer Rev (5' =&gt; 3')</th>
<th>Ann T C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Novel Isoforms: Exon Skipping</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES2</td>
<td>018640_1257_0905</td>
<td>GCTGTGGAGGAAGGGA</td>
<td>TAGGGATCGAGGCCTTTGC</td>
<td>57</td>
</tr>
<tr>
<td>ES3</td>
<td>057480_0973_1085</td>
<td>GCTGTAGGGGAAGTGGCTA</td>
<td>TGTGGATCTCTGGATGGCT</td>
<td>57</td>
</tr>
<tr>
<td><strong>Novel Isoforms: New Splice Patterns</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AI3</td>
<td>018086_0726_1326</td>
<td>TGCATCACATCTGCATCGAG</td>
<td>GTGTCCTCTATACCTTAGCGG</td>
<td>55</td>
</tr>
<tr>
<td>AI4</td>
<td>023612_1809_1670</td>
<td>TCCTTCAAGCTATCTCCTCT</td>
<td>GTGGGGACGGAATGATTAACA</td>
<td>59</td>
</tr>
<tr>
<td>AI5</td>
<td>239323_0568_3279</td>
<td>CATGCTTGTCCACCCAGG</td>
<td>CTTGGGAGACACAGGACTC</td>
<td>59</td>
</tr>
<tr>
<td>AI6</td>
<td>017234_1267_3837</td>
<td>GGCACAAACAATGCTGTACC</td>
<td>GTTATGCAGACACCTTTCTG</td>
<td>55</td>
</tr>
<tr>
<td>AI7</td>
<td>082654_1071_3200</td>
<td>ATCTCTACAAGATCTCCC</td>
<td>CCCACCTGATGTCATTC</td>
<td>55</td>
</tr>
<tr>
<td>AI8</td>
<td>285069_1182_1741</td>
<td>ACTGATTAAGCTAACCCTCAT</td>
<td>ATTTCTGGGCTGGAGCTA</td>
<td>55</td>
</tr>
<tr>
<td>AI9</td>
<td>011394_0819_2283</td>
<td>TGTGGTTAGAGCTAAAGATCT</td>
<td>TCTCATTGGCTATCTTGTGTA</td>
<td>55</td>
</tr>
<tr>
<td><strong>Novel Transcripts: Unspliced Intrinsic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UI1</td>
<td>075828_1754_1488</td>
<td>ACAACTCATCTGTGCTCAGC</td>
<td>GGACACAGGTAAAGACTCTC</td>
<td>55</td>
</tr>
<tr>
<td><strong>Novel Transcripts: Intergenic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INT1</td>
<td>006542_0430_2670</td>
<td>TCCTCCTTTTACCTTGGT</td>
<td>GGGAAATGTTACAAATACTGTG</td>
<td>51</td>
</tr>
<tr>
<td><strong>Fusion transcripts (F), genome deletion (D) and rare isoform (I)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4A (F)</td>
<td>107781_1044_1738</td>
<td>ACCTTCCCCTGGCTGAAGAG</td>
<td>CCAATAGCGGCAGGTTAA</td>
<td>58</td>
</tr>
<tr>
<td>1B (D)</td>
<td>167378_1645_3303</td>
<td>GCAGAGGTTGTACATCTCCC</td>
<td>CTCCAGCATCTTAACTTTTC</td>
<td>55</td>
</tr>
<tr>
<td>6B (I)</td>
<td>045624_1590_1179</td>
<td>AGCCTGGGCTGACTCAGCTAG</td>
<td>CGATGAGGCTGACTCTATTG</td>
<td>55</td>
</tr>
</tbody>
</table>
Experimental validation of selected transcript categories

<table>
<thead>
<tr>
<th></th>
<th>A17</th>
<th>A18</th>
<th>A19</th>
<th>4A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

Results:
- A17: OK n.d. n.d.
- A18: OK OK OK
- 4A: OK n.d. n.d.

1: 1360
2: 1345
3: 1645
Exploring the ncRNA world: identifying known ncRNAs

<table>
<thead>
<tr>
<th>ncRNA class</th>
<th>Number of unique ncRNAs matching the breast cancer library¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small</strong></td>
<td></td>
</tr>
<tr>
<td>snoRNA</td>
<td>32</td>
</tr>
<tr>
<td>piRNAs</td>
<td>6</td>
</tr>
<tr>
<td>scAluRNAs</td>
<td>23</td>
</tr>
<tr>
<td>snRNAs</td>
<td>1</td>
</tr>
<tr>
<td>snmRNAs</td>
<td>1</td>
</tr>
<tr>
<td><strong>Long regulatory RNAs²</strong>:</td>
<td></td>
</tr>
<tr>
<td>Host genes</td>
<td>11</td>
</tr>
<tr>
<td>Imprinted transcripts</td>
<td>4</td>
</tr>
<tr>
<td>Antisense transcripts</td>
<td>9</td>
</tr>
<tr>
<td>Cancer associated transcripts</td>
<td>11</td>
</tr>
<tr>
<td><strong>TUF</strong></td>
<td>26</td>
</tr>
<tr>
<td><strong>Expressed pseudogenes</strong></td>
<td>11</td>
</tr>
<tr>
<td><strong>Predicted conserved secondary structure³</strong></td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>96</td>
</tr>
</tbody>
</table>

²Both miRNA and snoRNA host transcripts are considered.
³According to RNASearch predictions (Torarinsson et al., 2006).
⁴Known ncRNAs were retrieved from both RNAdb and NONCODE databases (see Methods).

Abbreviations: snoRNA, small nucleolar RNAs; piRNA, Piwi-associated small RNAs; scAluRNA, small cytoplasmic Alu-repeat RNAs; snRNAs, small nuclear RNAs; snmRNAs, small mitochondrial RNAs; TUF, transcripts of unknown function.
Exploring the ncRNA world: identifying novel ncRNAs

UCSC human genome screenshot of a “gene desert” region (8 kb) on the X chromosome identified as transcribed by the sequencing read 002770_3171_2414. The read overlaps a CRITICA-predicted putative noncoding transcript (CR621898) and reveals a new, highly conserved transcriptional island, according to a vertebrate 28 multi-species alignment and PhastCons conservation score.
MALAT1 is a conserved 8-kb ncRNA whose expression correlates with the risk of developing metastasis in non-small-cell lung cancer (NSCLC) patients. We found 309 reads mapping along this regulatory ncRNAs, which, when assembled with the cap3 program, gave rise to 14 contigs distributed on all the length of the ncRNA. This suggests that MALAT-1 may be abundantly expressed in the analyzed breast cancer sample, in accordance with previous results.
An interesting ncRNA: MALAT-1

A detailed meta-analysis of publicly available gene expression datasets from the CleanEx database (http://www.cleanex.isb-sib.ch, Praz 2004) revealed nine breast cancer gene expression datasets suitable for MALAT1 expression analysis. Four of these datasets were composed of large, well characterized and stratified cohorts of patients and included control and Tamoxifen treatment regimes (GSE6532B, GSE4922B, GSE3494B, GSE1456B). When the normalized intensity of the probes corresponding to MALAT1 was analyzed we detected variable expression, but also observed a significant density of intensities above the normalized mean value. We selected a subset of 137 ER+ breast cancer patients (untreated) from the GSE6532B dataset (Loi 2007) and identified the data points derived from the 10 Affymetrix HG-U133B probesets corresponding to MALAT1. Unsupervised clustering revealed a wide range of expression of this transcript in this experiment, with good agreement between expression values from probesets in different samples, as shown in Fig. 1.
An interesting ncRNA: MALAT-1

MALAT1 distribution - LOI untreated dataset

log2 normalized intensity

number of samples

-2 -1 0 1 2 3

0 2 4 6 8 10 12

MALAT1 distribution - LOI Tam-treated dataset

log2 normalized intensity

number of samples

-2 -1 0 1 2 3

0 2 4 6 8 10 12
Summary of the findings

We classified 132,113 sequence reads mapping to the human genome in well-defined categories (intragenic, extragenic, novel transcripts, known and novel exons and isoforms) and detected a range of unusual transcriptional events that could be related to the disease, including one possible cancer-associated splice variant, seven putative gene fusions, two genic deletions and a novel isoform.

We validated by direct sequencing on the cDNA library two gene fusions, one deletion, one intergenic and one intragenic new transcript, two exon skipping isoforms and seven isoforms due to different splice site usage.

We also explored the non-protein-coding portion of the breast cancer transcriptome, identifying hundreds of copies of non-coding RNAs usually associated with other cancer types such as *MALAT1*, and many novel non-coding transcripts, supported by EST evidence and conservation analysis.

Institute of Biomedical Technologies, National Research Council, Milan, Italy
Department of Biochemistry and Molecular Biology, University of California Los Angeles, USA
Department of Biology and Genetics for Medical Sciences, University of Milan, Italy
ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia QLD 4072, Australia
Faculty of Pharmacological Sciences, University of Milan, Italy
Department of Biochemistry and Molecular Biology, University of Bari, Italy
Division of Pathology and Laboratory Medicine, European Institute of Oncology, Milan, Italy
Science and Technology Pole, Istituto di Ricovero e Cura a Carattere Scientifico MultiMedica, Milan, Italy
Translational Research Unit, Department of Experimental Oncology, Istituto Nazionale Tumori, Milan, Italy

HPC support: Ivan Merelli (ITB-CNR, Milan), Vital-IT team (SIB, Lausanne), CILEA staff (Segrate, Milan).
Statistics: Elia Biganzoli, dept. of Biostatistics, University of Milano.
Cariplo Foundation, FP6 and MIUR financed this project.

alessandro.guffanti@genomnia.com