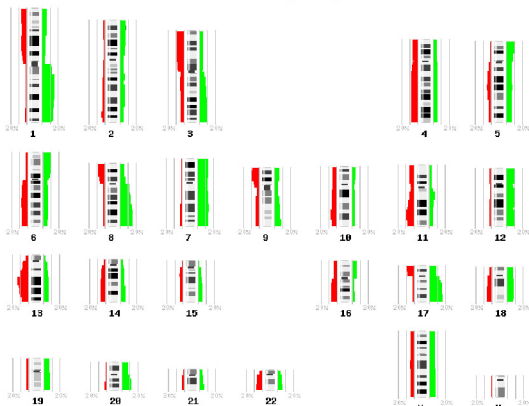


Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

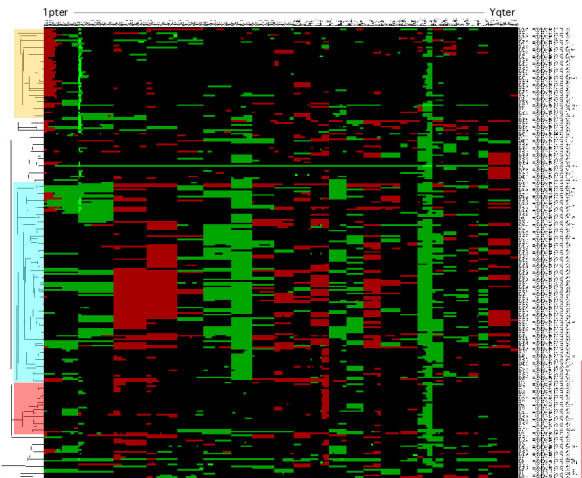
michael baudis, dept. of pathology, stanford university

Although the deciphering of the human genome has been pushed forward over the last years, little effort has been made to collect and integrate the treasure trove of clinical tumor cases analyzed by molecular-cytogenetic methods into current data schemes. Publicly announced at BCATS 2001, since then [progenetix.net] has been established as the largest public source of chromosomal imbalance data with band-specific resolution. Targets for the use of the data collection may be the description of prediction of oncogene and suppressor gene loci, identification of related loci for pathway creation, and especially the combination of the data with expression array experiments for filtering of relevant genes among the deregulated candidates.

Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosomal gains (green) and losses (red).

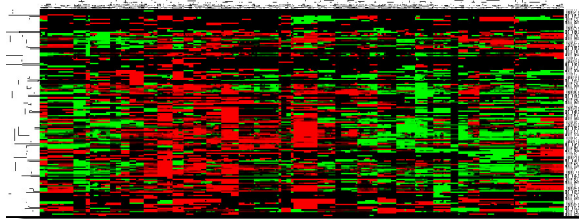


Material and Methods Chromosomal aberration data of more than 5478 cases from 196 publications describing results of Comparative Genomic Hybridization (CGH) experiments were collected. Minimal requirements were diagnosis of a malignant or benign neoplasia, analysis of clinical tumor samples and report of the analysis results on a case by case basis, resolved to the level of single chromosomal bands. Data was transformed from the diverse annotation formats to standardized ISCN "rev ish" nomenclature. For the transformation of the non-linear ISCN data to a two-dimensional matrix with code for the aberration status of each chromosomal band per case, a reverse pattern matching algorithm was developed in Perl. Graphical representations and cluster images are generated for all different subsets (Publications, ICD-O-3 entities, meta-groups) and presented on the progenetix.net website.



Clustering of the band averages for the different ICD-O entities

Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (10 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (ave. 23.3% with max. 96.2% at 3p14p26) and pheochromocytomas (ave. 10.9% with max. 92.7% at 3p); prominent gain maxima were found in pure high grade infiltrating duct carcinomas of the breast (ave. 5.9% with max. 95.7% at 11q13), T-PLL (ave. 4.7% with max. 81.8% for whole 8q) and dedifferentiated liposarcomas (ave. 10.4% with max. 81.8% at 12q13), among others.

By cluster analysis, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsummed in the same diagnostic entity; the example of neuroblastomas is shown.

Examples of hotspots of genomic imbalance

SCLC, pheochromocytoma, high grade DCIS, T-PLL, dedifferentiated liposarcoma



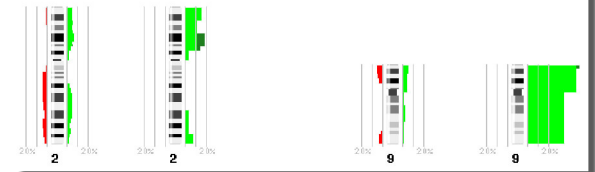
Conclusion So far, progenetix.net project was able to:

1. collect a large dataset of genomic aberration data generated through a molecular-cytogenetic screening technique (CGH)
2. develop the software tools to transform those data to a meta format compatible to commonly used genomic interval descriptions
3. produce graphical and numerical output from those data for hot spot detection and statistical analysis.

For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and possibly for the description of common and divergent genetic pathways in the oncogenetic process of different tumor entities. The transformed raw data of the progenetix.net collection is available for research purposes over the website.

Distinction of histologically related through their chromosomal aberration pattern

Amplification of the REL locus on 2p16 and gain of 9p(ter) distinguishes primary mediastinal B-cell lymphomas (PMBL, right) from diffuse large cell lymphomas (DLCL, left). The distinction may have clinical implications



Identification of different aberration patterns in Neuroblastoma (289 cases)

N-Myc (2p25) amplification is the hallmark of a subgroup, showing only constant loss of the terminal portion 1p. Other groups are defined by the loss of parts of 11q, or a "chromosomal instability" phenotype. Gains on 17q are a common feature of all groups.

Those patterns may be combined with gene-level information to reconstruct the different pathways leading to malignant transformation.