Next Generation (Sequencing) Tools for

Nucleotide-Based Information



GENome-wide eXpression PROfiling



Massive Analysis of cDNA Ends for simultaneous Genotyping and Transcription Profiling in High Throughput

Björn Rotter, PhD GenXPro GmbH, Frankfurt am Main

www.genxpro.de



- I: What we do at GenXPro
- II: Massive Analysis of cDNA Ends for "TranSNiPtomics"

II.a: Reliable SNP detection II.b: Highly accurate Transcript-Quantification II.c: Bioinformatics

III: Example of Workflow for "TranSNiPtomics"

Our Service Portfolio

Nucleotide-based information

Transcripts :

eXpression PROfiling

- RNAseq
- SuperSAGE, MACE
- Real-Time qPCR service
- Normalization of cDNA libraries
- microRNA

Genomics:

Epigenomics:

- Genetic Markers (RAD, RC-Seq)
- copy number variations
- de novo sequencing (mate pair, Reduced Redundancy)

Methylation-specific DK (MSDK) MethSeq

Bioinformatics: - NGS Data Handling, Assembly, Quantification, BLAST - Expression-Data Interpretation: Functional analysis



"TranSNiPtomics" =

simultaneous analysis of gene expression AND polymorphisms; advantages:

- Markers located within genes very likely connected to specific trait
- Markes can be chosen from differentially expressed genes to increase chance of involvement in trait

Requirements:

- Sufficient coverage distinguish between sequencing error and SNP
- Accurate measurement of transcription levels





Most of the transcript species are expressed at low levels (below 10 copies per million). Frequent transcripts make up 50 % of all transcripts.



One solution: RNA-Seq

Average Transcript Size: 2 500 bp



Mapping/Assembly, Quantification



But:

for medium and low-level transcripts, e.g. of transcription factors or receptors, very (very...) deep sequencing is required in order to distinguish a SNP from a sequencing error with RNAseq (low coverage per basepair).

Our solution = MACE: only the cDNA-3'ends are sequenced

- concentration on the most polymorphic site in a gene !
- highly specific for good annotation !
- easy to quantify !
- low costs !



Massive Analysis of cDNA Ends (MACE):



GENome-wide eXpression PROfiling



Fragmentation



GENome-wide eXpression PROfiling

Ge



2nd generation sequencing of 50-100 bp

Ger

GENome-wide

eXpression PROfiling





Massive Analysis of cDNA Ends: MACE

How it works

eXpression PROfiling

GENome-wide

Assembly & Counting













Assembly & Counting

Ger

GENome-wide

eXpression PROfiling





Counting, BLAST

Only one fragment per transcript!

TranSNiPtomics- why MACE?

High coverage to distingish between SNP and error

MACE



Concentration on polymorph 3' end: SNPs with enough coverage : 2

GENome-wide

eXpression PROfiling



Reads distributed all over transcript: SNPs with enough coverage : 0



GENome-wide

eXpression PROfiling

Coverage for SNP detection

Wheat, nucleosome/chromatin assembly factor C; 160 TPM

MACE, 20 Mio Reads

L	V	1	1	т	V		N	S		L	1		Y	V		E	L	A		R	L	1	/	К	L	М	R V		Y	V	н
I	AGT	GC	ATA	CT	GT	A A	A C	A G	CT	TA	AT	AT	ATC	STO	GG	A A C	TT	GC	G A	GA	CT	T G T	TA	A A	CT	GATG	AGAGT	GT	ATG	TGC	A
			10200		Institute																							a			
t.	04 10 04		1		í.					1		1			1		1		- (T.	1		í.	20 - 104 	253 113	53 00 09		1	267 112	67
T	A G T	GC	ATA	СТ	GT	A A	AC	A G	СТ	TA	AT	ATA	ATC	STO	GG	AAC	TT	GC	0 4		СТ	T G T	C A	AA	A	200 02	33 CV105	-		207 02	
	A A G G T	0 0				A A A A A A A A A A A A A A A A A A A										а на				A A <th></th> <th></th> <th>Т Т А А А А А А А А А А А А А А А А А А</th> <th>. A A A A A A A A A A A A A A A A A A A</th> <th></th> <th>G A T G G A T G G G A T G G A G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G G A T G G A T G G A T G G A T G G A T G G A T</th> <th>A A</th> <th></th> <th>A T T G</th> <th>T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C C T T G C C C C T T G C C C C C T G C C</th> <th>AAAG AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA</th>			Т Т А А А А А А А А А А А А А А А А А А	. A A A A A A A A A A A A A A A A A A A		G A T G G A T G G G A T G G A G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G A T G G G A T G G A T G G A T G G A T G G A T G G A T	A A		A T T G	T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C T T G C C C T T G C C C C T T G C C C C C T G C C	AAAG AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA



RNA seq, 20 Mio reads, same position



Coverage too low!



- I: What we do at GenXPro's
- II: Massive Analysis of cDNA Ends for "TranSNiPtomics"

II.a: Reliable SNP detection II.b: Highly accurate Transcript-Quantification II.c: Bioinformatics

III: Example of Workflow for "TranSNiPtomics"



For the same depth of analysis, RNA-Seq requires about 20-30 times more sequencing*

*Asmann et. al 2009



"TrueQuant" Technology solves problem of PCR-introduced bias

Certain tags or fragments are preferentially amplified during PCR



The Solution: GenXPro's bias-proof "TrueQuant" technology:

→ PCR-based copies are eliminated from dataset



Problem of PCR-introduced BIAS

GENome-wide eXpression PROfiling

TrueQuant-corrected vs. uncorrected Data



Negative common logarithm of the p-value for differential expression of gene expression comparisons (Audic & Claverie; 1997). From human pancreatic tumors.



Pearson correlation coefficient = 0.9983357 (!)





- I: What we do at GenXPro's
- II: Massive Analysis of cDNA Ends for "TranSNiPtomics"

II.a: Reliable SNP detection II.b: Highly accurate Transcript-Quantification II.c: Bioinformatics

III: Example of Workflow for "TranSNiPtomics"





GENome-wide

eXpression PROfiling

Example: Picea abies MACE analysis, 10 Mio Tags (from Igor Yakovlev)

Example: all differentially expressed transcripts belonging to GO term: "cytosol"(in red)



Massive Analysis of cDNA ends: MACE

SNP-Analysis

SNP

 $\overline{1}$

Prephenate dehydratase, alleles

	ACT GT <mark>R</mark> CTT C <mark>RR GTR</mark> C	CRRGCTTCGGCCCCCT	C GGCT GG <mark>R</mark> T GRR GTT	G <mark>CT GCT GTRR</mark> T GGTT GRR	RTT GGAR GTARRTARRTRT GT GTT
	ACT GT <mark>R</mark> CTT C <mark>RR</mark> GT <mark>R</mark> C	C <mark>RR</mark> GCTTCGGCCCCCT	CGGCTGG <mark>RTGRRGTT</mark>	A CT G CT GT AAT G GTT GAA	I <mark>ATT</mark> GG <mark>RA</mark> G <mark>TARAT</mark> ARAAT <mark>AT</mark> GT GTT
	- CT GT <mark>R</mark> CTT C <mark>RR</mark> GT <mark>R</mark> C	C <mark>RR</mark> GCTTCGGCCCCCT	C GGCT GG <mark>R</mark> T GRR GTT	G <mark>CT GCT GTRR</mark> T GGTT GRP	<mark>RTT</mark> GG <mark>AR</mark> G <mark>TARATARATAT</mark> GGG <mark>TTT</mark>
	- CT GT <mark>R</mark> CTT C <mark>RR</mark> GT <mark>R</mark> C	C <mark>RR</mark> GCTTCGGCCCCCT	C G G C T G G R T G R R G T 1	G <mark>CT GCT GTRR</mark> T GGTT GRR	L <mark>RTT</mark> GG <mark>RR GTRRR TRRRTTRG GTTT</mark>
	- CT GT <mark>R</mark> CTT C <mark>RR GTR</mark> C	C <mark>RR</mark> GCTTCGGCCCCCT	C G G C T G G R T G R R G T T	G <mark>CT GCT GTRR</mark> T GGTT GRR	IRTT GGRR GTRARTRARTRT GT GTTT
	CTTCARGTRC	C <mark>RR</mark> GCTTCGGCCCCCT	C G G CT G G RT G R R G T T	G <mark>CT GCT GTRR</mark> T GGTT GRR	<mark>ATT GGRA GTAARARAR T</mark> AT GT GTTTATAT C -
	CTT CRAGTAC	C <mark>RRGCTTCGG</mark> CCCCCT	C G G CT G G AT G A A G T T	GCT GCT GTRRT GGTT GRA	ATT GGAR GTARATARATAT GT GTTTATAT C –
	CTT CAR GTA C	C <mark>RRGCTTCGG</mark> CCCCCT	CGGCTGGRTGRRGTT	GCT GCT GTRRT GGTT GRP	RTT GGRR GTRARTRARTAT GT GGTTRTRT C -
	CTT CARGTAC	CRAGCTT CGGCCCCCT	C G G C T G G R T G R R G T 1	RET GET GTRRT GGTT GRR	ATT GGAR GTARATARATAT GT GTTTATAT C -
	RCTTCRRGTRC	CRR GCTT CGGCCCCCT	CGGCT GGRT GRR GTT	R CT G CT GTRRT G G TT GRR	ATT GGAA GTAAATAAATAT GT GTTTATAT
	RCTTCARGTRC	CRAGCTTCGGCCCCCT	CGGCTGGRTGRAGT	GCT GCT GTRAT GGTT GRA	ATT GGAA GTAAATAAATAT GT GTTTATAC
	TRCTTCARGTRC	CRAGCTTCGGCCCCCT	CGGCTGGRTGRRGTT	A CT GCT GTAAT GGTT GAA	ATT GGAR GTARATARATAT GT GT TTATA
	GTACTTCAAGTAC	CRAGCTTCGGCCCCCT	CGGCTGGATGAAGT	A CT G CT GT A A T G G T T G A A	RTTGGRAGGRARTRARTATGTGTTTAT
	TGTACTTCAAGTAC	CARGETTEGGECCCCT	CGGCTGGATGAAGTT	a et get gtaat ggtt gaa	ATT GGAA GTAAATAAATAT GT GT TTA
	TGTACTTCAAGTAC	CAAGUTTUGGUUUUUT	CGGCTGGATGAAGT	A CTGCTGTAATGGTTGAA	ATT GGAA GTAAATAAATAT GT GT TTA
	CT GTACTT CAAGTAC	CAAGCTTCGGCCCCCT	CGGCTGGATGAAGT	A DT GCT GTAAT GGTT GAA	ATT GGAA GTAAATAATAT GT GT TT
CTTT C C C	CIGINGINGIIGANGING	CARGUITUGGUUUUUT	CCCCTCCTTCTTCTT	NET CET CTART GGII GAR	PTTCCPPCTPPPT
crere e	CT CTR CTT CRR CTR C	CARGOTTOGGCCCCCCT	CCCCTCCPTCPP CPP CT	NET CET CT DE T CETT CEL	
GTTTTTTTTCTTCTTTTTTTCCCCTTTCCCC	CTCTD CTTCDD CTDC	CARGOTTCGGCCCCCCT	CCCCTCCTTCTTCTTCTT	CTCCTCTPPTCCTTCPP	PTTCCPPCTPPPTPTPTCTCTCTTPTPTPTP
TTTT GGGTTTGCG	CTGTR CTTCRRGTRC	CREGCTTCGGCCCCCCT	CGGCTGGRTGRRGT	GTTGCTGTRATGGTTGRA	BTTT CBB CCBB
BTTTT GGGTTT GCG	CT GTR CTT CRAGTAC	CRAGCTTCGGCCCCCT	CGGCTGGATGAAGT	GET GET GTAAT GETT GAA	RTTGGRRGGR
GTTCATTAATTTT GGGTTTGCG	CT GTR CTT CRAGTAC	CARGCTTCGGCCCCCT	CGGCTGGATGAAGTT	A CT G CT GTAAT G GTT GAA	RT
TTTTGTTCATTAATTTTGGGTTTGCG	ACT GTACTT CAAGTAC	CRAGCTTCGGCCCCCT	CGGCTGGRTGRRGTT	ACT GCT GTAAT GGTT G	
TTTTTTTGTTCATTAATTTTGGGTTTGCG	CT GTRCTT CRAGTRC	CRRGCTTCGGCCCCCT	C G G CT G G RT G R R G T 1	RET GET GTRRT GG	
GTTTTTTTT GTT C <mark>R</mark> TT <mark>RR</mark> TTTT GGGTTT GC G	ACT GTACTT CAAGTAC	CRAGCTTCGGCCCCCT	C GGCT GGAT GAA GTT	GCT GCT GTRRT	
GTTTTTTTTGTTC <mark>R</mark> TT <mark>RR</mark> TTTTGGGGTTTGCG	ACT GTRCTT CARGTRC	CRAGCTTCGGCCCCCT	CGGCTGGRTGRRGT1	G <mark>CT GCT GTR</mark> CT	

ClustalX view of MACE-sequences

GENome-wide

eXpression PROfiling

Ger

GenXPro

eXpression PROfiling

Massive Analysis of cDNA Ends: MACE

Advantages

GENome-wide

High-resolution, exact quantification of expressed genes:

- only one specific tag is sequenced per transcript
- even rare transcripts are analyzed

High SNP-coverage for gene-specific markers:

- only the highly polymorphic 3' ends are sequenced
- high coverage for reliable SNP detection, even in rare transcripts

High throughput & low costs:

- hundreds of samples can be analyzed simultaneously
- multiplexed sequencing reduces costs while still providing high resolution and coverage for SNPs
- Ideally suited for large sets of genotypes



"Simultaneous detection of SNPs in differentially expressed transcripts is now affordable in hundreds of genotypes."

Including

Sample preparation: no antisense artefacts Sequencing: ~10 Mio Reads Annotation: optimized annotation routine Quantification: no PCR bias "TrueQuant" Gene Expression Analysis; Enriched GO Terms or Pathways

No bioinformatics required, no costly software required, no hardware required!



