Introduction to Bioinformatics

Introduction to NGS/Genomics Technologies (GSEA)

2021.

A preview of the past few weeks

Profiles are lists of quantified molecular features

A preview of the past few weeks

Profiles are lists of quantified molecular features

There are profiles of …

- RNA Transcripts (mRNA, miRNA, lncRNA, ...)
- Proteins (total expression, phosphorylation, ubiquitination …)
- Metabolites (intra cellular, secreted, …)
- Epigenetics (DNA methylation, histone methylation, histone acetylation)
- Transcription factor binding (ChIP)
- DNA copy number variation
- Microbiomes (16S rRNA, Metagenomes, …)

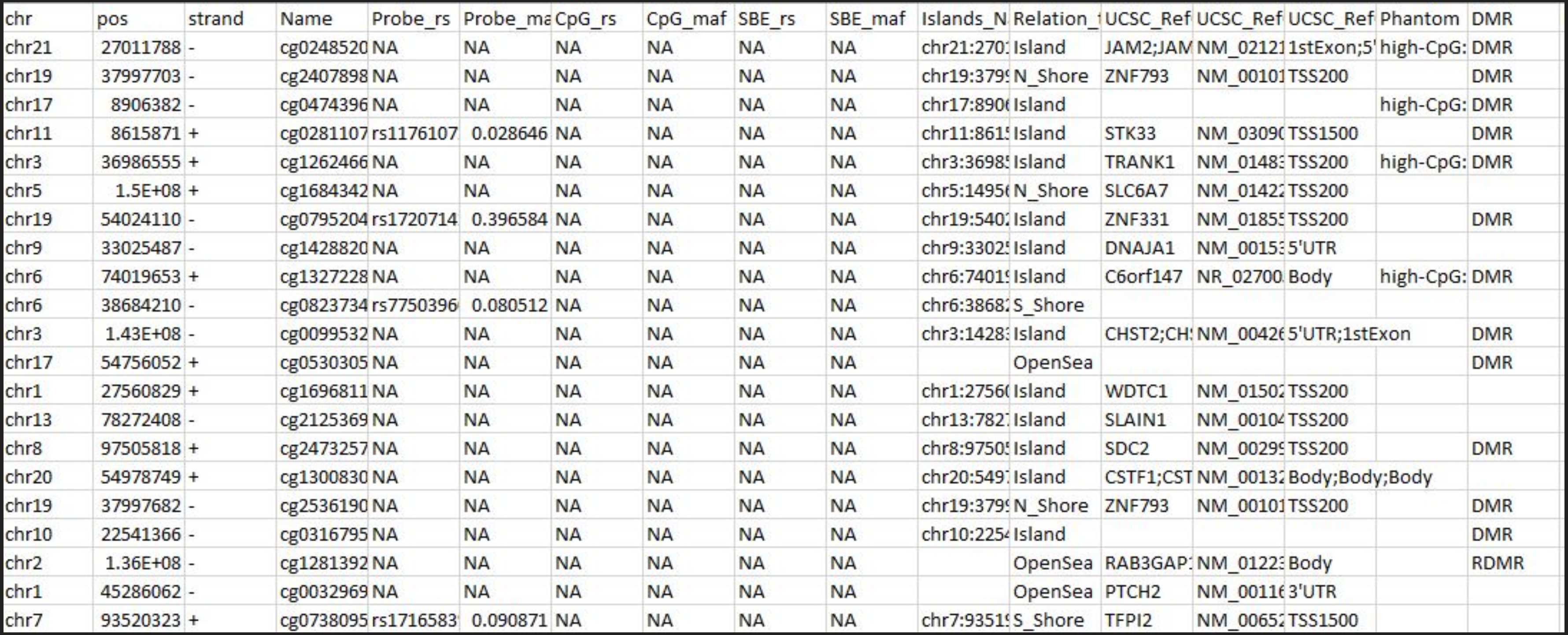
A preview of the past few weeks

Profiles are lists of quantified molecular features

Profiles can be generated by different technologies

- RNA Transcripts (microarray, nanoString, RNAseq)
- Proteins (MassSpec, protein array)
- Metabolites (NMR, MassSpec,...)
- Epigenetics (ChIP-seq, bisulfate sequencing, ATAC-seq)
- Transcription factor binding (ChIP)
- DNA copy number variation (aCGH, NGS)
- Immune cell infiltration (FACS, imaging, proteomics)
- Microbiomes (arrays, 16S rRNA-seq)

We got this Excel now what?



chr	pos	strand	Name	Probe_rs	Probe_ma	CpG_rs	CpG_maf	SBE_rs	SBE_maf	Islands_N	Relation_	UCSC_Ref	UCSC_Ref	UCSC_Ref	Phantom	DMR
chr21	27011788	-	cg0248520	NA	NA	NA	NA	NA	NA	chr21:270	Island	JAM2;JAM	NM_02121	1stExon;5'	high-CpG:	DMR
chr19	37997703	-	cg2407898	NA	NA	NA	NA	NA	NA	chr19:379	N_Shore	ZNF793	NM_00101	TSS200		DMR
chr17	8906382	-	cg0474396	NA	NA	NA	NA	NA	NA	chr17:890	Island				high-CpG:	DMR
chr11	8615871	+	cg0281107	rs1176107	0.028646	NA	NA	NA	NA	chr11:861	Island	STK33	NM_0309	TSS1500		DMR
chr3	36986555	+	cg1262466	NA	NA	NA	NA	NA	NA	chr3:3698	Island	TRANK1	NM_0148	TSS200	high-CpG:	DMR
chr5	1.5E+08	+	cg1684342	NA	NA	NA	NA	NA	NA	chr5:1495	N_Shore	SLC6A7	NM_0142	TSS200		
chr19	54024110	-	cg0795204	rs1720714	0.396584	NA	NA	NA	NA	chr19:540	Island	ZNF331	NM_0185	TSS200		DMR
chr9	33025487	-	cg1428820	NA	NA	NA	NA	NA	NA	chr9:3302	Island	DNAJA1	NM_0015	5'UTR		
chr6	74019653	+	cg1327228	NA	NA	NA	NA	NA	NA	chr6:7401	Island	C6orf147	NR_02700	Body	high-CpG:	DMR
chr6	38684210	-	cg0823734	rs7750396	0.080512	NA	NA	NA	NA	chr6:3868	S_Shore					
chr3	1.43E+08	-	cg0099532	NA	NA	NA	NA	NA	NA	chr3:1428	Island	CHST2;CH	NM_0042	5'UTR;1stExon		DMR
chr17	54756052	+	cg0530305	NA	NA	NA	NA	NA	NA		OpenSea					DMR
chr1	27560829	+	cg1696811	NA	NA	NA	NA	NA	NA	chr1:2756	Island	WDTC1	NM_0150	TSS200		
chr13	78272408	-	cg2125369	NA	NA	NA	NA	NA	NA	chr13:782	Island	SLAIN1	NM_0010	TSS200		
chr8	97505818	+	cg2473257	NA	NA	NA	NA	NA	NA	chr8:9750	Island	SDC2	NM_0029	TSS200		DMR
chr20	54978749	+	cg1300830	NA	NA	NA	NA	NA	NA	chr20:549	Island	CSTF1;CST	NM_0013	Body;Body;Body		
chr19	37997682	-	cg2536190	NA	NA	NA	NA	NA	NA	chr19:379	N_Shore	ZNF793	NM_00101	TSS200		DMR
chr10	22541366	-	cg0316795	NA	NA	NA	NA	NA	NA	chr10:225	Island					DMR
chr2	1.36E+08	-	cg1281392	NA	NA	NA	NA	NA	NA		OpenSea	RAB3GAP	NM_0122	Body		RDMR
chr1	45286062	-	cg0032969	NA	NA	NA	NA	NA	NA		OpenSea	PTCH2	NM_0011	3'UTR		
chr7	93520323	+	cg0738095	rs1716583	0.090871	NA	NA	NA	NA	chr7:9351	S_Shore	TFPI2	NM_0065	TSS1500		

Gene Set Enrichment Analysis

- Gene set enrichment analysis (GSEA) is a statistical method to determine if predefined sets of genes are differentially expressed in different phenotypes

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Gene Set Enrichment Analysis

- Gene set enrichment analysis (GSEA) is a statistical method to determine if predefined sets of genes are differentially expressed in different phenotypes
- Predefined gene sets may be genes in a known metabolic pathway, located in the same cytogenetic band, sharing the same Gene Ontology category, or any user-defined set

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Gene Set Enrichment Analysis

- Gene set enrichment analysis (GSEA) is a statistical method to determine if predefined sets of genes are differentially expressed in different phenotypes
- Predefined gene sets may be genes in a known metabolic pathway, located in the same cytogenetic band, sharing the same Gene Ontology category, or any user-defined set
- In array experiments where no single gene shows statistically significant differential expression between phenotypes, GSEA has identified significant differentially expressed sets of genes

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Gene Set Enrichment Analysis

- Gene set enrichment analysis (GSEA) is a statistical method to determine if predefined sets of genes are differentially expressed in different phenotypes
- Predefined gene sets may be genes in a known metabolic pathway, located in the same cytogenetic band, sharing the same Gene Ontology category, or any user-defined set
- In array experiments where no single gene shows statistically significant differential expression between phenotypes, GSEA has identified significant differentially expressed sets of genes
- GSEA is likely to be more powerful than conventional single-gene methods for studying the large number of common diseases in which many genes each make subtle contributions

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Why GSEA?

- The conventional statistical analysis method for array experiments is to
  - examine one gene at a time,
  - determine a p-value that the gene is differentially expressed/methylated in different phenotypes
  - apply a correction (penalty) to the p-value for having tested multiple genes (described further below)

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Why GSEA?

- The conventional statistical analysis method for array experiments is to
  - examine one gene at a time,
  - determine a p-value that the gene is differentially expressed/methylated in different phenotypes
  - apply a correction (penalty) to the p-value for having tested multiple genes (described further below)

- This method had limitations
  - One such that; In common diseases, a large number of genes each make subtle contributions, and these genes are difficult to detect in single gene analyses.

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.

Why GSEA?

- The conventional statistical analysis method for array experiments is to
  - examine one gene at a time,
  - determine a p-value that the gene is differentially expressed/methylated in different phenotypes
  - apply a correction (penalty) to the p-value for having tested multiple genes (described further below)

- This method had limitations
  - One such that; In common diseases, a large number of genes each make subtle contributions, and these genes are difficult to detect in single gene analyses.

Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. Jing Shi et al.